Machine learning a zpracování dat pomocí Microsoft Azure

Bc. Lukáš Beran

Diplomová práce
2015

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky
akademický rok: 2014/2015

ZADÁNÍ DIPLOMOVÉ PRÁCE
(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bc. Lukáš Beran**
Osobní číslo:	**A13777**
Studijní program:	**N3902 Inženýrská informatika**
Studijní obor:	**Automatické řízení a informatika**
Forma studia:	**kombinovaná**
Téma práce:	**Machine learning a zpracování dat pomocí Microsoft Azure**
Téma anglicky:	**Machine Learning and Data Processing Using Microsoft Azure**

Zásady pro vypracování:

1. **Seznamte se s oblastí strojového učení (Machine Learning).**
2. **Vytvořte praktické návody použití nástroje Microsoft Azure v oblasti Machine Learning.**
3. **Navrhněte vhodnou strukturu dat a vhodná reálná data pro použití s Microsoft Azure.**
4. **Implementujte ukázkovou praktickou aplikaci.**
5. **Proveďte analýzu výsledného řešení.**

Rozsah diplomové práce:

Rozsah příloh:

Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam odborné literatury:

1. **BARGA, Roger, Wee Hyong TOK a FONTAMA. Predictive Analytics with Microsoft Azure Machine Learning: Build and Deploy Actionable Solutions in Minutes. 1. vyd. English: Apress, 2014. ISBN 978-1484204467.**
2. **BISHOP, Christopher M. Pattern recognition and machine learning. New York: Springer, c2006, xx, 738 s. ISBN 0-387-31073-8.**
3. **MITCHELL, Tom M. Machine learning. Boston: WCB/McGraw-Hill, c1997, xvii, 414 s. ISBN 0-07-042807-7.**
4. **ABU-MOSTAFA, Yaser S, Malik MAGDON-ISMAIL a Hsuan-Tien LIN. Learning from data: a short course. Pasadena, CA?: AML Book, c2012, xii, 201 s. ISBN 978-1-60049-006-4.**
5. **ALPAYDIN, Ethem. Introduction to machine learning. 2nd ed. Cambridge, Massachusetts: MIT Press, c2010, xl, 537 s. ISBN 978-0-262-01243-0.**
6. **Microsoft's Cloud Platform: Azure [online]. Seattle, Dostupné z: http://azure.microsoft.com/en-us/**
7. **ŠENOVSKÝ, Pavel. Modelování rozhodovacích procesů, skripta, 2. vydání. Ostrava: Vysoká škola báňská Technická univerzita Ostrava, 2009.**
8. **BÍLA J.: Umělá inteligence a neuronové sítě v aplikacích, ČVUT, 1996, ISBN 80-01-01275-1.**

Vedoucí diplomové práce: **doc. Ing. Zuzana Komínková Oplatková, Ph.D.**
Ústav informatiky a umělé inteligence

Datum zadání diplomové práce: **27. února 2015**

Termín odevzdání diplomové práce: **20. května 2015**

Ve Zlíně dne 27. února 2015

L.S.

doc. Mgr. Milan Adámek, Ph.D.
děkan

prof. Ing. Vladimír Vašek, CSc.
ředitel ústavu

Prohlašuji, že

- beru na vědomí, že odevzdáním diplomové/bakalářské práce souhlasím se zveřejněním své práce podle zákona č. 111/1998 Sb. o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších právních předpisů, bez ohledu na výsledek obhajoby;
- beru na vědomí, že diplomová/bakalářská práce bude uložena v elektronické podobě v univerzitním informačním systému dostupná k prezenčnímu nahlédnutí, že jeden výtisk diplomové/bakalářské práce bude uložen v příruční knihovně Fakulty aplikované informatiky Univerzity Tomáše Bati ve Zlíně a jeden výtisk bude uložen u vedoucího práce;
- byl/a jsem seznámen/a s tím, že na moji diplomovou/bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, zejm. § 35 odst. 3;
- beru na vědomí, že podle § 60 odst. 1 autorského zákona má UTB ve Zlíně právo na uzavření licenční smlouvy o užití školního díla v rozsahu § 12 odst. 4 autorského zákona;
- beru na vědomí, že podle § 60 odst. 2 a 3 autorského zákona mohu užít své dílo – diplomovou/bakalářskou práci nebo poskytnout licenci k jejímu využití jen připouští-li tak licenční smlouva uzavřená mezi mnou a Univerzitou Tomáše Bati ve Zlíně s tím, že vyrovnání případného přiměřeného příspěvku na úhradu nákladů, které byly Univerzitou Tomáše Bati ve Zlíně na vytvoření díla vynaloženy (až do jejich skutečné výše) bude rovněž předmětem této licenční smlouvy;
- beru na vědomí, že pokud bylo k vypracování diplomové/bakalářské práce využito softwaru poskytnutého Univerzitou Tomáše Bati ve Zlíně nebo jinými subjekty pouze ke studijním a výzkumným účelům (tedy pouze k nekomerčnímu využití), nelze výsledky diplomové/bakalářské práce využít ke komerčním účelům;
- beru na vědomí, že pokud je výstupem diplomové/bakalářské práce jakýkoliv softwarový produkt, považují se za součást práce rovněž i zdrojové kódy, popř. soubory, ze kterých se projekt skládá. Neodevzdání této součásti může být důvodem k neobhájení práce.

Prohlašuji,

- že jsem na diplomové práci pracoval samostatně a použitou literaturu jsem citoval. V případě publikace výsledků budu uveden jako spoluautor.
- že odevzdaná verze diplomové práce a verze elektronická nahraná do IS/STAG jsou totožné.

Ve Zlíně

podpis diplomanta

ABSTRAKT

Tato diplomová práce se zabývá možnostmi Machine Learning v Microsoft Azure. V teoretické části práce je nahlédnuto do historie strojového učení v Microsoftu, popsány jsou praktické příklady využití strojového učení a součástí jsou i dvě případové studie využití Azure Machine Learning v praxi, z nichž jedna popisuje inteligentní řízení univerzitní budovy. V praktické části práce jsou názorně ukázány možnosti využití Azure Machine Learning na předpovědi hodnocení filmů.

Klíčová slova: strojové učení, azure, microsoft, zpracování dat

ABSTRACT

This Master's thesis deals with the possibilities of Machine Learning in Microsoft Azure. In the theoretical part of the thesis is looked into the history of machine learning in Microsoft, described are specific examples of using machine learning and included are two case studies of the use of Azure Machine Learning in practice, one of which describes the intelligent management of a university building. In the practical part of the thesis are clearly presented how to use Azure Machine Learning predictions on movie ratings.

Keywords: machine learning, azure, microsoft, data processing

Děkuji vedoucí mé diplomové práce paní doc. Ing. Zuzaně Komínkové Oplatkové, Ph.D. za velmi cenné připomínky na konzultacích a vedení mé práce. Dále děkuji společnosti Microsoft za poskytnutí bezplatného přístupu do Microsoft Azure, jmenovitě pak paní Daniele Liškové a panu Brandonu Beckovi za vstřícnost a poskytnuté informace. A samozřejmě děkuji i své rodině a přítelkyni za trpělivost a podporu.

Prohlašuji, že odevzdaná verze bakalářské/diplomové práce a verze elektronická nahraná do IS/STAG jsou totožné.

OBSAH

ÚVOD **8**
I. TEORETICKÁ ČÁST **9**
1 STROJOVÉ UČENÍ **10**
2 VŠUDYPŘÍTOMNÉ DOPORUČOVÁNÍ **12**
3 MICROSOFT AZURE **16**
4 SLUŽBY MICROSOFT AZURE **17**
4.1 ACTIVE DIRECTORY **17**
4.2 BACKUP **17**
4.3 CDN **17**
4.4 HDINSIGHT **17**
4.5 MACHINE LEARNING **17**
4.6 SITE RECOVERY **18**
4.7 STORAGE **18**
4.8 VIRTUAL MACHINES **18**
4.9 WEBSITES **18**
5 AZURE MACHINE LEARNING **19**
5.1 HISTORIE STROJOVÉHO UČENÍ VE SPOLEČNOSTI MICROSOFT **19**
5.2 PŘÍKLAD STROJOVÉHO UČENÍ V MICROSOFT MALWARE PROTECTION CENTER **20**
5.3 PŘÍPADOVÉ STUDIE MACHINE LEARNING **21**
5.3.1 JJ FOOD SERVICE 21
5.3.2 CARNEGIE MELLON UNIVERSITY 22
5.4 MACHINE LEARNING STUDIO **22**
5.5 SLUŽBA AZURE ML API **24**
5.6 CENY AZURE MACHINE LEARNING **24**
II. PRAKTICKÁ ČÁST **27**
6 MICROSOFT AZURE MACHINE LEARNING V PRAXI **28**
6.1 PRVNÍ KONTAKT S AZURE MACHINE LEARNING **28**
6.2 VYTVOŘENÍ VLASTNÍHO EXPERIMENTU **35**
6.2.1 ZÍSKÁNÍ DAT 35
6.2.2 PŘEDZPRACOVÁNÍ DAT 38
6.2.3 DEFINOVÁNÍ PARAMETRŮ 43
6.2.4 VÝBĚR A APLIKACE UČÍCÍHO ALGORITMU 44
6.2.5 PŘEDPOVĚDI NAD NOVÝMI DATY 47
7 ANALÝZA DATABÁZE FILMŮ A PREDIKCE HODNOCENÍ **50**
7.1 DATA DOSTUPNÁ Z IMDB **50**
7.2 TVORBA MODELU V MICROSOFT AZURE **51**
7.3 PŘEDZPRACOVÁNÍ DAT **53**
7.4 POROVNÁNÍ MODELŮ PŘEDPOVĚDI **56**

7.4.1 Studie 1: Boosted Decision Tree, parametr průměrných herců 56
7.4.2 Studie 2: Boosted Decision Tree, všechny parametry 56
7.4.3 Studie 3: Neural Network Regression, parametr průměrných herců 56
7.4.4 Studie 4: Neural Network Regression, všechny parametry 57
7.4.5 Studie 5: Decision Forest Regression, parametr průměrných herců 57
7.4.6 Studie 6: Decision Forest Regression, všechny parametry 57
7.4.7 Studie 7: Linear Regression, parametr průměrných herců 58
7.4.8 Studie 8: Linear Regression, všechny parametry 58
7.4.9 Studie 9: Boosted Decision Tree, podstatná korelace 58
7.4.10 Studie 10: Boosted Decision Tree, podstatná korelace, experimentální nastavení 59
7.4.11 Studie 11: Boosted Decision Tree, podstatná korelace, dostatek hodnot 59
7.4.12 Studie 12: Boosted Decision Tree, experimentálně zjištěné parametry i nastavení 60
7.4.13 Studie 13: Boosted Decision Tree, experimentálně zjištěné parametry, nastavení dle Sweep Parameters 60
7.4.14 Studie 14: Boosted Decision Tree, experimentálně zjištěné parametry, nastavení dle Sweep Parameters, více než 30 hodnocení 61
7.5 ZHODNOCENÍ VÝSLEDKŮ 63
ZÁVĚR 64
SEZNAM POUŽITÉ LITERATURY 65
SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK 70
SEZNAM OBRÁZKŮ 71
SEZNAM TABULEK 73

ÚVOD

Strojové učení je v současné době velmi populární téma, jelikož požadavky na předpovědi a doporučování se týkají velké části komerční i nekomerční sféry.

Microsoft Azure je cloudová platforma, která poskytuje velké množství služeb pro firmy i jednotlivce. Machine Learning patří k nejnovějším službám dostupným v Microsoft Azure a nabízí techniky strojového učení, které svou jednoduchostí a kvalitně zpracovanou dokumentací uspokojí jak běžné uživatele se základní znalostí statistiky a strojového učení, tak i náročné velké firmy, které mohou využít podporu jazyka R nebo Python, garantovanou dostupnost a technickou podporou.

Cílem práce je vytvoření praktického návodu pro práci s Microsoft Azure Machine Learning a praktická ukázka možností na vlastním příkladu.

Využití strojového učení se nachází i v oblasti automatického řízení, například pro inteligentní řízení budov za účelem snižování nákladů na provoz a zvyšování komfortu.

Tato práce se zabývá možnostmi využití Azure Machine Learning pro oblast strojového učení. Praktický příklad využití je ukázán na předpovědi hodnocení filmů.

Téma práce jsem si vybral z důvodu, že mne oblast strojového učení a umělé inteligence velmi zaujala v průběhu studia a v Microsoft Azure Machine Learning vidím velký potenciál pro využití v mnoha oblastech.

I. TEORETICKÁ ČÁST

1 STROJOVÉ UČENÍ

Strojové učení je technika, která prožívá v poslední době velký rozmach. Jednoduše řečeno, strojové učení převádí datasety do částí software, které jsou známé jako modely, které mohou reprezentovat tyto datasety, zobecňovat je a pomocí nich tvořit predikce nových dat. Tyto techniky využívají například vyhledávače, ať už ty globální obecného typu Google nebo Bing, tak i specializované vyhledávače například v internetových obchodech. Součástí toho pak mohou být i doporučení, která se zobrazují návštěvníkům. Kupříkladu z historie nákupů na internetovém obchodě s elektronikou může být patrné, že zákazníci, kteří si koupí digitální fotoaparát, si k němu koupí i ochranné pouzdro. Proto při nákupu fotoaparátu se zákazníkovi může objevit doporučení, že by si ke svému fotoaparátu mohl koupit i ochranné pouzdro. Stejným způsobem může fungovat například i cílení reklamy, kde pokud si ve svém oblíbeném vyhledávači vyhledáme nějaký produkt, v cílené reklamě se nám pak i na jiných webech s reklamním systémem společnosti provozující daný vyhledávač může zobrazit reklama na internetový obchod, který prodává produkt, který jsme dříve vyhledávali.

Strojové učení ale nemusí být součástí jen marketingu nebo prodeje, ale na strojovém učení a heuristikách jsou založeny například i bezpečnostní produkty, jako jsou antiviry, filtry spamu apod., které se stále učí nová pravidla pro třídění a rozpoznávání. Používá se také pro detekci anomálií nebo chyb, tvorbu předpovědí nebo pro stále populárnější virtuální asistenty jako jsou Cortana od společnosti Microsoft [12] nebo Siri od konkurenční firmy Apple [13].

Výše uvedené příklady jsou ale jen zlomkem toho, kde všude se strojové učení používá. Strojové učení můžeme rozdělit na tři základní skupiny:

1. **Data Mining** (česky dolování z dat), kde strojové učení slouží pro získání poznatků a závislostí z rozsáhlých databází.
2. **Statistical Engineering** (česky statistické techniky), kde strojové učení slouží pro převod dat na software, který poté může tvořit rozhodovací mechanismy nad neúplnými daty.
3. **Artificial Intelligence** (česky umělá inteligence), kde strojové učení slouží pro emulaci lidského myšlení, díky čemuž mohou počítače „vidět, slyšet a chápat“.

Strojové učení ale není jednoduchá technika. Obvykle vyžaduje složitý software, výkonné počítače a zkušené vědce, kteří dané problematice rozumí a dokáží s ní správně pracovat. Před spuštěním projektu na strojové učení je nutné udělat hloubkovou expertízu. K tomu

všemu jsou zapotřebí odborníci na analýzu dat, kterých je ovšem nedostatek, a jsou velmi drazí. Z tohoto důvodu může být pro firmy nebo jednotlivce velmi nevýhodné budovat si strojové učení vlastními silami, ale mnohem výhodnější, rychlejší a jednoduší může být pronajmout si strojové učení jako službu. A na tomto principu funguje i cloudová služba [14] Microsoft Azure Machine Learning (Azure ML), která zpřístupňuje strojové učení i širokému spektru potenciálních uživatelů.

2 VŠUDYPŘÍTOMNÉ DOPORUČOVÁNÍ

Kvalitní doporučovací systémy jsou potřebné všude kolem nás. Když si vybíráme film nebo knihu, případně firmy hledají potenciální zaměstnance na sociálních sítích jako LinkedIn, zde všude oceníme doporučení od nějakého automatického systému.

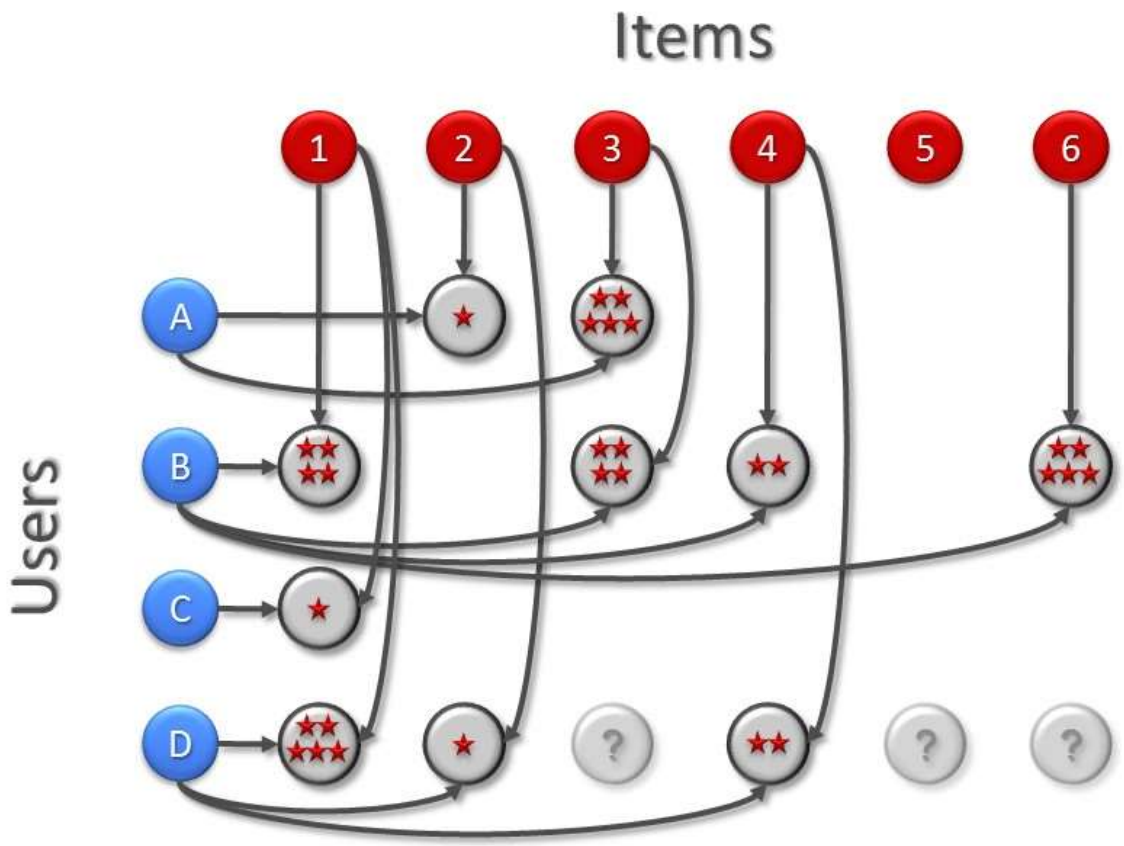
V systémech doporučení typicky existují dvě základní entity. Pro základní příklad je pojmenujme jako *uživatele* a *položky*. Uživatelé jsou lidé, kterým chceme něco doporučit. Položky jsou věci, které budeme danému uživateli doporučovat. Může se jednat například o výše zmíněné knihy, filmy nebo jiné lidi.

Budeme chtít tedy například uživateli doporučit nějakou dobře hodnocenou restauraci. Doporučení tedy rozdělíme do dvou základních kroků:

1. Předpověď, jak by daný uživatel hodnotil každou z vhodných restaurací (například v jeho okolí nebo podle typu restaurace).
2. Ze seznamu těchto vhodných restaurací vybereme tu, kterou by uživatel ohodnotil nejlépe podle naší předpovědi z prvního bodu. Tuto restauraci mu doporučíme.

Jenže jak předpovědět, jak by uživatel ohodnotil všechny restaurace, když je nikdy nehodnotil? Zde přichází na scénu strojové učení.

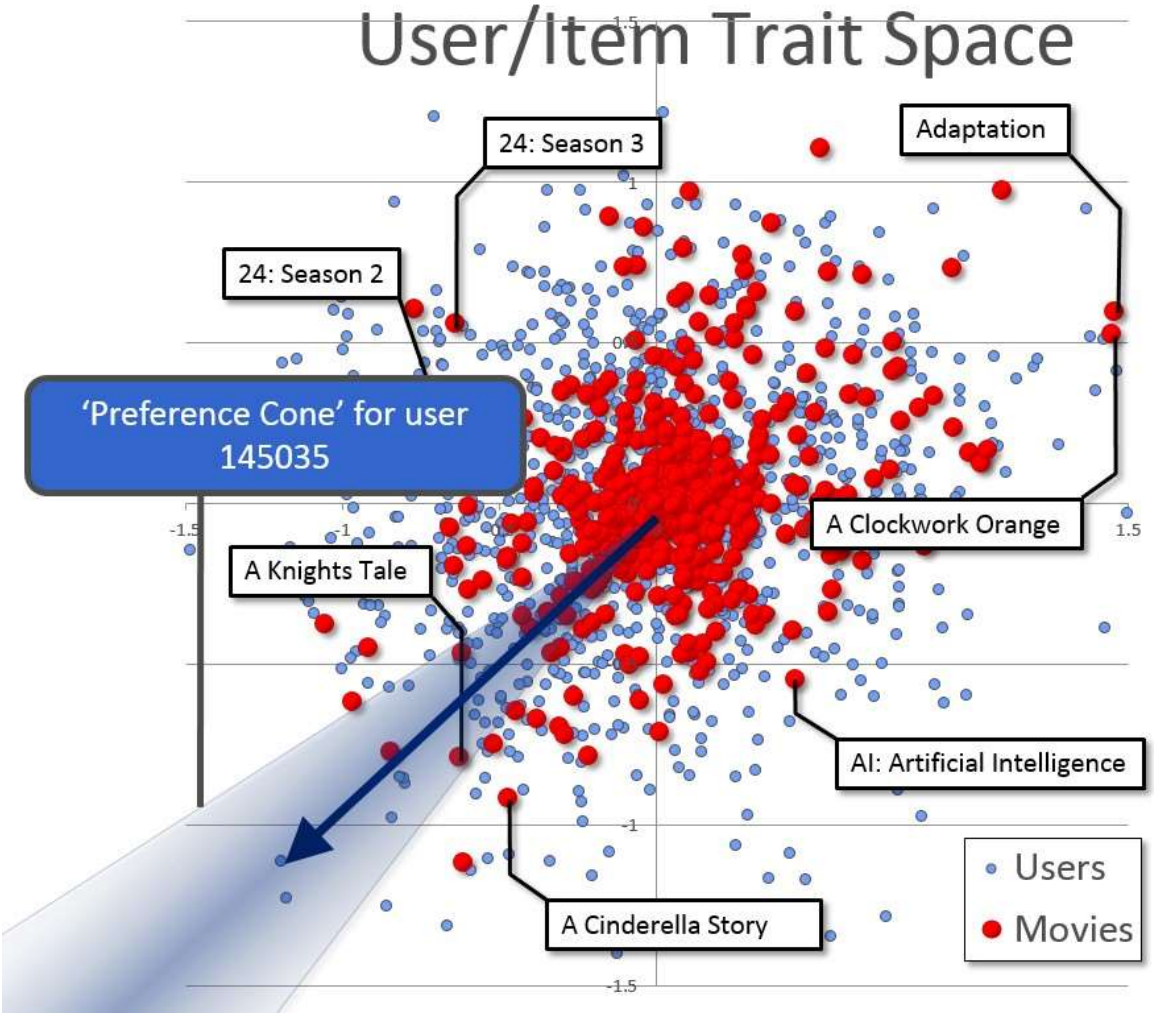
Při budování modelu pro strojové učení, které bude předpovídat hodnocení restaurací, budeme potřebovat nasbírat data o uživatelích, položkách (restauracích) a hodnoceních. Můžeme si to představit jako matici na obrázku (Obrázek 1), kde uživatelé jsou řádky, položky jsou sloupce a záznamy jsou hodnocení.



Obrázek 1 - Matice uživatelů, položek a hodnocení. Převzato z [5]

Daná matice bude řídká, protože uživatelé ohodnotí jen velmi malý počet položek. Systém strojového učení poté jako výstup jednoduše vrátí funkci, která předpovídá, jak by daný uživatel ohodnotil danou položku. Předpověď hodnocení však nemusí být jediným rozhodujícím faktorem pro doporučení. Jiné informace, jako například nákupy, prokliky nebo čas strávený na stránce, mohou být ekvivalentní nebo i důležitější pro výsledné doporučení.

Jak tedy takové doporučení funguje? Systém doporučení se učí vkládání uživatelů a položek do tzv. latentního prostoru parametrů, viz obrázek (Obrázek 2). Modré tečky jsou uživatelé, kteří hodnotí položky znázorněné červenými tečkami pozitivně, pokud jsou jejich vektory zarovnány s vektory položek. A naopak pokud jsou vektory opačné, hodnotí uživatelé dané položky negativně. Podobní uživatelé a podobné položky budou umístěny blíže k sobě společně v prostoru parametrů, což umožní odvodit hodnocení kombinací uživatel/položka, u kterých není hodnocení dostupné z trénovacích dat.



Obrázek 2 - Dvojrozměrný latentní prostor parametrů. Převzato z [5]

Na obrázku výše (Obrázek 2) je vidět dvojrozměrný prostor parametrů pro jeho jednoduchou vizualizaci. V praxi se však používají dvaceti až padesáti rozměrové prostory.

Občas je možné najít obecně interpretovatelné parametry na některých osách grafu. Například na obrázku (Obrázek 2) směr „sever-jih“ může být „dospělí-děti“.

Klíčovým problémem systému doporučení je start. Noví uživatelé nemusí mít dostatečné množství hodnocení a nové položky nemusí mít hodnocení od dostatečného množství uživatelů, aby bylo možné je považovat za dostatečně reprezentativní vzorek pro tvorbu předpovědí. Proto systémy strojového učení (včetně Azure Machine Learning) umožňují uživatele a položky reprezentovat jako vektory pomocí dalších metadat[1]. U uživatelů to mohou

[1] Strukturovaná data o datech. Umožňují blíže specifikovat danou věc pomocí dodatečných informací.

být věk, pohlaví či geolokace uživatele. U filmů to může být žánr, délka, rok vydání, herci apod. Tato data pak mohou sloužit pro přesnější doporučení. [5]

3 MICROSOFT AZURE

Microsoft Azure je cloudová služba společnosti Microsoft, kterou je možné využít při budování infrastrukturních služeb (IaaS[2]) nebo platformních služeb (PaaS[3]). Microsoft Azure je globálně dostupná služba provozovaná v několika desítkách datacentrech po celém světě na všech kontinentech, díky čemuž dosahuje vysoké dostupnosti (99,9 % až 99,95 % podle typu použitých služeb). Samozřejmostí je nonstop technická podpora. [1], [2]

Díky vlastnostem cloudu je možné služby Microsoft Azure velmi jednoduše škálovat. Vhodná je tedy jak pro malé projekty s nízkými rozpočty, tak i pro obrovské projekty s vysokými nároky na výkon i dostupnost. Jak už to u tohoto typu služeb bývá, platí se pouze za skutečně využité prostředky a parametry služeb je možné měnit podle potřeby nebo si přidělování prostředků naplánovat například s ohledem na pracovní dobu firmy, díky čemuž lze významně ušetřit náklady na provoz.

Microsoft Azure lze také kombinovat do tzv. hybridních cloudů, ve kterých část infrastruktury běží ve vlastním datovém centru (privátní cloud) a část běží u externího dodavatele (veřejný cloud), přičemž oba cloudy spolu úzce spolupracují a pro koncového uživatele se tváří jako jeden ucelený systém.

[2] Infrastructure as a Service, jedna z možností distribuce služeb v cloud computingu, kdy poskytovatel nabízí výpočetní infrastrukturu v dojednané konfiguraci (servery, datová úložiště, síťové prvky a další) jako službu zákazníkům za pravidelný poplatek.

[3] Platform as a Service, oproti IaaS se tento typ distribuce služeb liší v tom, že poskytovatel zajišťuje i operační systém celého řešení včetně potřebných nadstaveb, díky čemuž zákazníkovi odpadá nutnost starat se o nasazení, správu a aktualizaci software.

4 SLUŽBY MICROSOFT AZURE

Microsoft Azure obsahuje velké množství služeb a pravidelně přibývají další nové. Aktuálně (22. září 2014) je dostupných 31 služeb. V následujícím textu popíši několik nejvýznamnějších služeb.

4.1 Active Directory

Služba Azure Active Directory poskytuje správu identit a přístupových oprávnění. Identity je možné synchronizovat s lokálním řešením a aktivovat jednotné přihlašování do více služeb.

4.2 Backup

Azure Backup spravuje cloudové zálohy pomocí základních nástrojů ve Windows Server nebo System Center.

4.3 CDN[4]

Azure CDN umožňuje doručit datově náročný obsah koncovým uživatelům po celém světě s nízkou latencí a vysokou rychlostí díky robustní síti globálních datacenter.

4.4 HDInsight

Azure HDInsight Service je služba založená na frameworku Apache™ Hadoop® [31]. Díky vysokému výkonu cloudu umožňuje rychle a efektivně zpracovávat velká data (Big Data) bez nutnosti pořízení vlastního výkonného a drahého hardware.

4.5 Machine Learning

Azure Machine Learning umožňuje jednoduše navrhovat, testovat, operacionalizovat a spravovat prediktivní analytická řešení v cloudu. Služba Azure Machine Learning umožňuje přímé napojení na HDInsight pro řešení s velkými daty. Azure Machine Learningu se budu blíže věnovat ve své práci.

[4] Content Delivery Network, velký distribuovaný systém serverů určený pro doručování datově náročného obsahu uživatelům po celém světě.

4.6 Site Recovery

Azure Site Recovery pomáhá firmám chránit dostupnost svých služeb pomocí koordinace replikace a obnovy privátních cloudů.

4.7 Storage

Azure Storage poskytuje geograficky redundantní datové úložiště skrze klasické SMB[5] sdílení.

4.8 Virtual Machines

Azure Virtual Machines umožňuje nasazení Windows Server nebo Linux virtuálních strojů v cloudu.

4.9 Websites

Azure Websites slouží pro nasazení webových aplikací na škálovatelné cloudové infrastruktuře.

Microsoft Azure obsahuje velké množství služeb a pravidelně přibývají další nové [1], [2].

[5] Server Message Block, síťový komunikační protokol aplikační vrstvy sloužící ke sdílenému přístupu k souborům, tiskárnám a další komunikaci mezi uzly v síti.

5 AZURE MACHINE LEARNING

Azure Machine Learning je výkonná cloudová prediktivní analýza, která je určená jak pro běžné uživatele, tak i pro zkušené odborníky. Využívá ověřené techniky, které používají například vyhledávač Bing, herní konzole Xbox nebo laboratoře Microsoft Research.

5.1 Historie strojového učení ve společnosti Microsoft

I když se to mnoha lidem nemusí zdát, Microsoft má více než 20 let zkušeností se strojovým učením, což je mnohem více, než jak dlouho známe pojmy jako Big Data [15] nebo Deep Learning [16]. Tato skutečnost dává Microsoftu a jeho Azure Machine Learningu velkou konkurenční výhodu.

První principy strojového učení se v Microsoftu datují do roku 1992, kdy začali pracovat s Bayesian Networks na modelování přirozené řeči a rozpoznávání hlasu. Bayesovská síť je grafický model, ve kterém jsou zakódované pravděpodobnostní vztahy mezi proměnnými. Oproti statistickým technikám má však několik výhod. Díky tomu, že kóduje závislosti všech proměnných, snadno zvládá situace, kdy chybí některé záznamy. Bayesovské sítě také mohou být použity pro učení kauzálních vztahů, a proto mohou být použity pro pochopení základního problému a předpovídání důsledků iterací. [6] V 90. letech také zjistili, že mnoho problémů, jako jsou kategorizace textu nebo priorizace e-mailu, jsou řešitelné pomocí kombinace lineární klasifikace a Bayesových sítí. Lineární klasifikace je rozhodovací problém, který pomocí lineární kombinace charakteristik objektu přiřadí danému objektu skupinu nebo třídu, do které daný objekt patří. Výsledkem byl první detektor spamu založený na analýze obsahu.

S postupem času se vyvíjené algoritmy stávaly více sofistikovanými a mohly být použity pro řešení problémů spojených se strojovým učením, jako jsou například počítačové vidění a širší uplatnění rozpoznávání hlasu. Rozhodovací stromy byly použity například při provádění pixel-wise klasifikací[6] pro odhad lidské pózy, kterou používá například senzor Kinect, nebo pro analýzu obrazů v medicíně, například u skenů z tomografu, kde tyto algoritmy vyvinuté v laboratořích Microsoftu byly v roce 2012 schváleny organizací FDA pro detekci a lokalizaci u počítačové tomografie. Rozhodovací stromy jsou jedním z nejjednodušších

[6] Klasifikace pomocí vlastností jednotlivých pixelů obrazu (intenzita, barva, spektrální informace, …).

nástrojů pro podporu rozhodování. Jedná se o orientované grafy, které nám v rozhodování pomáhají vizualizací všech možných řešení problému do větví stromu. Pro všechny tyto možnosti poté vypočítáme užitnost a jejím porovnáním se rozhodneme pro nejvhodnější řešení. Rozhodování poté probíhá buď v režimu jistoty, nebo nejistoty, přičemž nejistota je v praktickém použití častější. U jistoty můžeme předem říct, jaké bude mít rozhodnutí následky. U nejistoty tyto následky nejsme schopni přesně stanovit, a proto kromě předpokládaných následků kalkulujeme i s pravděpodobností. Pro rozhodování v režimu jistoty se používají deterministické stromy, pro nejistotu jsou stromy nedeterministické (stochastické). [7]

Mezi nepoužívanější frameworky[7] pro strojové učení dlouhodobě patří frameworky klasifikace a regrese. V těchto frameworcích se strojové učení učí mapování z vektoru dat do štítku (klasifikace) nebo do hodnoty (regrese). Štítky a hodnoty ale nejsou tím jediným, co ze strojového učení vzniká. Jako jedním z prvních příkladů bylo učení se hodnosti (learning to rank), kde výstupem je ohodnocený seznam prvků, který je velmi vhodný například pro vyhledávač Bing. Dalším populárním frameworkem je konstrukce kauzálních modelů (construction of causal models), který se používá pro modelování reklamních systémů.

5.2 Příklad strojového učení v Microsoft Malware Protection Center

Aktuálně se v byznysu kolem antimalware[8] ochrany používají tři zdroje signálů strojového učení:

1. Dobrovolně zvolené telemetrické údaje o zjištěných hrozbách.
2. Vlastní analýza škodlivých souborů.
3. Informace od partnerů.

Každý měsíc systémy strojového učení v Microsoft Malware Protection Center zanalyzují více než 30 milionů různých souborů. Tyto soubory porovnávají se známými soubory, webovými stránkami a způsoby používání a podle toho vytváří a nasazují podpisy pro soubory identifikované jako škodlivé. Díky velkému množství analyzovaných dat je možné rychle

[7] Framework je softwarová struktura sloužící jako podpora při programování.
[8] Malware je škodlivý software, který má za cíl poškodit operační systém, shromažďovat citlivé informace nebo získat přístup k cizímu počítači.

hledat a identifikovat nový malware. Automatická analýza je kontrolována, doplňována a upravována bezpečnostními odborníky, díky čemuž je ochrana chytřejší a přesnější.

Kontrola pomocí databází malware je doplněna analýzou komunikace, kdy podezřelá komunikace na síti může být blokována ještě dřív, než dojde k aktualizaci databáze škodlivého software. [4]

5.3 Případové studie Machine Learning

I když je služba Machine Learning teprve v testovací fázi, využívá ji již mnoho firem a organizací v reálném nasazení.

5.3.1 JJ Food Service

JJ Food Service je jedna z největších britských firem zabývajících se doručováním jídla. Firma s více než 60 000 zákazníky a 4 500 produkty v nabídce potřebovala zajistit, aby vytvoření objednávky bylo pro zákazníka co nejrychlejší a nejjednodušší. Za roky fungování měli obrovské množství dat o zákaznících, čehož chtěli využít.

Mezi jejich zákazníky patří velké restaurace, které nechtějí trávit čas procházením zdlouhavých nabídek dodavatelů, ale na druhou stranu mají relativně předvídatelné potřeby. Taková restaurace například bude chtít každý den čerstvou zeleninu, mouku jednou za dva týdny a fritovací olej jednou za měsíc. A přesně takto by to měl jejich e-shop primárně nabízet bez nutnosti ručně upravovat nabídky pro každého zákazníka, ale vše automaticky na pozadí s tím, jak se vyvíjí potřeby zákazníků. K tomu použili data nasbíraná za poslední tři roky pro trénování datového modelu a Azure ML napojili na jejich webový systém a do call centra. Přitom celý proces zabral tři měsíce.

Celý systém se ale neustále vyvíjí a postupně přibývají další funkcionality. Například monitorování stavu nákupního košíku během nakupování a na základě toho doporučení dalšího zboží, které zákazník mohl zapomenout vložit nebo by mohl chtít.

Takto doporučené zboží tvoří aktuálně asi 5 % z objemu nákupů, což se nemusí zdát jako mnoho, ale při celkovém počtu objednávek se jedná o vysoké číslo v obratu firmy, ale také pro zákazníka to znamená přidaný komfort navíc, což ve výsledku může znamenat mnohem víc. [8]

5.3.2 Carnegie Mellon University

Carnegie Mellon University používá Microsoft Azure pro snížení nákladů na údržbu a provoz budov. Machine Learning používají pro detekci poruch, diagnostiku a efektivnější operace. Díky tomu se jim podařilo snížit spotřebu energií o 20 %. Požadavkem byla predikce v reálném čase, která by správcům umožnila například vyměnit nebo opravit opotřebované komponenty ještě před jejich selháním. Dalším požadavkem bylo zvýšení efektivity kontroly vytápění a chlazení díky lepšímu přizpůsobení nastavení jednotlivých termostatů. To vše muselo být rychlé, jednoduché na implementaci a přístupné i pro netechnické pracovníky.

Pro natrénování byla použita historická data z OSIsoft PI System. Výsledek se poté porovnával se skutečnými hodnotami z historie, aby se dosáhlo co nejpřesnějších předpovědí do budoucna.

Dalším cílem byla predikce teploty v budově, aby se mohlo lépe regulovat vytápění. Požadavkem byla teplota 72°F v budově od 9 hodin ráno. Před instalací Azure ML bylo vytápění nastaveno tak, že se spustilo v 6 hodin ráno, což přibližně odpovídalo požadavku. Vnitřní teplotu však ovlivňuje mnoho dalších faktorů, tudíž zde docházelo k plýtvání za vytápění. Proto nový systém postavený na Azure ML vyhodnocuje aktuální vnitřní i venkovní teplotu, očekávanou úroveň solární radiace a další faktory. Informace o solární radiaci nebyly ale dostupné, proto vědci museli tuto hodnotu nejdříve předpovídat. Vytrénovali model solární radiace pomocí algoritmu Boosted Decision Trees [17] dostupného v Azure ML, otestovali tento model na přesnost a poté použili v modelu vnitřní teploty. [9]

5.4 Machine Learning Studio

Velkou výhodou Azure Machine Learningu je jeho jednoduchost, protože umožňuje i lidem bez hlubších znalostí analýzy dat začít tvořit předpovědi. Machine Learning Studio, integrované vývojářské prostředí, používá k vytvoření experimentů gesta přetažení a jednoduché grafy toku dat. Díky tomu je možné mnoho úloh zadávat bez jediného řádku kódu. Kromě výše uvedeného ML Studio obsahuje i knihovnu ukázkových experimentů a algoritmů přímo z oddělení výzkumu Microsoftu.

Aby byl Azure Machine Learning konkurenceschopný a použitelný i pro odborníky zabývající se studiem dat, podporuje jazyk R [18], který je oblíbeným open source programovacím prostředím pro vytváření statistik a dolování dat. Kromě toho se dají experimenty i sdílet, takže již aplikované výzkumy mohou využít i další lidé.

Pro využití všech uvedených vlastností není potřeba instalace žádného software nebo konfigurace vlastního hardware. Vše je možné řešit přímo přes webový prohlížeč odkudkoliv na světě.

ML Studio hostuje velké množství modulů s grafickým rozhraním, jako jsou moduly pro data ingestion a transformace, dále pak moduly pro tvorbu, ohodnocení a vyhodnocení prediktivních modelů. Některé nejpokročilejší algoritmy jsou základem i pro Bing a Xbox.

Součástí jsou i škálovatelné open source machine learning balíky jako Vowpal Wabbit [19].

Jak již bylo zmíněno výše, podpora jazyka R je samozřejmostí. Je možné použít i vlastní již hotový R kód a zahrnout ho do experimentů ve spolupráci s dostupnými algoritmy. Podporováno je nyní více než 350 R balíčků a nové stále přibývají.

Součástí poskytovaných nejmodernějších algoritmů v ML Studiu jsou algoritmy jako škálovatelné posílené rozhodovací stromy (Scalable Boosted Decision Trees), Bayesovské doporučovací systémy (Bayesian Recommendation systems), hluboké neuronové sítě (Deep Neural Networks) [20] a rozhodovací džungle (Decision Jungles) [21] vyvinuté v Microsoft Research laboratořích.

ML Studio podporuje také machine learning algoritmy pro vícehodnotové [22] a binární klasifikace [23], regrese [24] a clustrování [25].

ML Studio podporuje i spolupráci více členů týmu.

Data mohou být do ML Studia dodána buď nahráním lokálních souborů, nebo importována pomocí modulu readeru. Lokální soubory mohou být nahrány jako datasety pomocí přidání nových datasetů v ML Studiu. Modul readeru může číst data z Azure Table [26], Azure Blob [27], SQL Database (Azure) [26] nebo HDInsight, případně pomocí http. ML Studio podporuje velikost datasetu až 10 GB. U Web Services není žádný limit velikosti datasetu. Podporováno je rozdělení pomocí Hive [28] nebo SQL dotazů [29]. V případě, že je potřeba pracovat s většími daty než 10 GB, je možné vytvořit více datasetů a použít moduly „Partition and Sample“, „Split“ nebo „Join“ pro rekombinaci těchto datasetů v ML Studiu na vytvoření trénovacích setů pro vytvoření prediktivních modulů.

U datasetů větších než pár GB je doporučeno nahrát data do Azure úložiště, SQL Database (Azure) nebo použít HDInsight místo klasického nahrávání z lokálního souboru.

5.5 Služba Azure ML API

Služba Azure ML API umožňuje nasazení prediktivních modelů vytvořených v ML Studiu jako škálovatelné webové služby odolné proti chybám. Webové služby vytvořené v ML API jsou REST API [30] poskytující rozhraní pro komunikaci mezi externími aplikacemi a vytvořeným prediktivním analytickým modelem. Webová služba poskytuje cestu pro komunikaci s prediktivním modelem v reálném čase pro přebírání výsledků predikcí a pro začlenění výsledků do různých externích klientských aplikací. Existují dva typy, a to Batch Execution Service (BES) pro asynchronní dávkový přístup a Request-Response Service (RRS) pro synchronní reakce s nízkou latencí.

Request-Response Service je webová služba s nízkou latencí určená k poskytování rozhraní pro bezstavové modely. Batch Execution Service je služba pro asynchronní ohodnocování dávky datových záznamů. Oba typy služeb mají podobné vstupy. Hlavní rozdíl je v tom, že BES čte blok dat z různých zdrojů jako bloby[9], tabulky v Azure, SQL Databáze (Azure), HDInsight (Hive Query) a http zdroje. Výsledkem ohodnocení je výstup do souboru v Azure blob úložišti a end-point je vrácen jako reakce.

Batch Execution Service je vhodná v případě, kdy chceme ohodnocovat velké množství datových bodů v dávkách nebo velká část dat je již ve formátu souboru v Azure úložišti nebo Hadoop clusteru. Webová služba může transformovat čtená data ještě před odesláním do modelu.

Request-Response Service je vhodná v případech, kdy je prediktivní analýza potřeba v téměř reálném čase.

Pro snadnou práci s ML API slouží generované stránky s nápovědou a ukázkami kódu v C#, R a Pythonu. [3]

5.6 Ceny Azure Machine Learning

Ceník platný pro finální globálně dostupnou verzi Azure Machine Learning je platný od 1. dubna 2015.

[9] Binary large object je označení pro datový typ blíže nespecifikovaných dat v databázi. Obvykle se jedná o obrázky, zvuky nebo jiná data.

Tabulka 1 - Ceny Azure Machine Learning

ML Seat Subscription	Měsíční poplatek	$9,99/Seat/měsíc
ML Studio	Hodinový poplatek	$1/hodinu výpočtu
ML API	Hodinový poplatek	$2/API hodinu výpočtu
	Poplatek za transakci	$0,50/1000 API transakcí

Machine Learning Seat reprezentuje workspace v Azure. Každý uživatel pracující s daným workspace by měl být pokrytý licencí ML Seat Subscription.

Azure Machine Learning je možný používat i ve zdarma dostupné verzi, která je však funkčně omezená.

Tabulka 2 - Porovnání parametrů služeb Azure Machine Learning

	Free	**Standard**
Ověřování	Microsoft Account	Azure Account
Maximální počet modulů v experimentu	100	Neomezeně
Maximální doba výpočtu experimentu	1 hodina	Neomezeně
Maximální úložiště	10 GB	Neomezeně
Výkon	Single Node	Multiple Nodes
Pracovní Web API	Ano (omezený výkon)	Ano (volitelný výkon)
Produkční Web API	Ne	Ano
SLA[10]	Ne	Ano

[10] Service-level agreement je smlouva definující parametry a kvalitu poskytované služby.

Rozdíl mezi Single Node a Multiple Nodes je v počtu současně zpracovávaných výpočtů. Zatímco Single Node umožňuje současně zpracovávat pouze jeden výpočet (jeden běžící modul), Multiple Nodes umožňuje současný běh neomezeného počtu výpočtů dle potřeby.

V rámci SLA je garantována 99,95% dostupnost API transakcí. Pro Batch Execution Service (BES) a management API je garance 99,9% dostupnost API transakcí. [35]

II. PRAKTICKÁ ČÁST

6 MICROSOFT AZURE MACHINE LEARNING V PRAXI

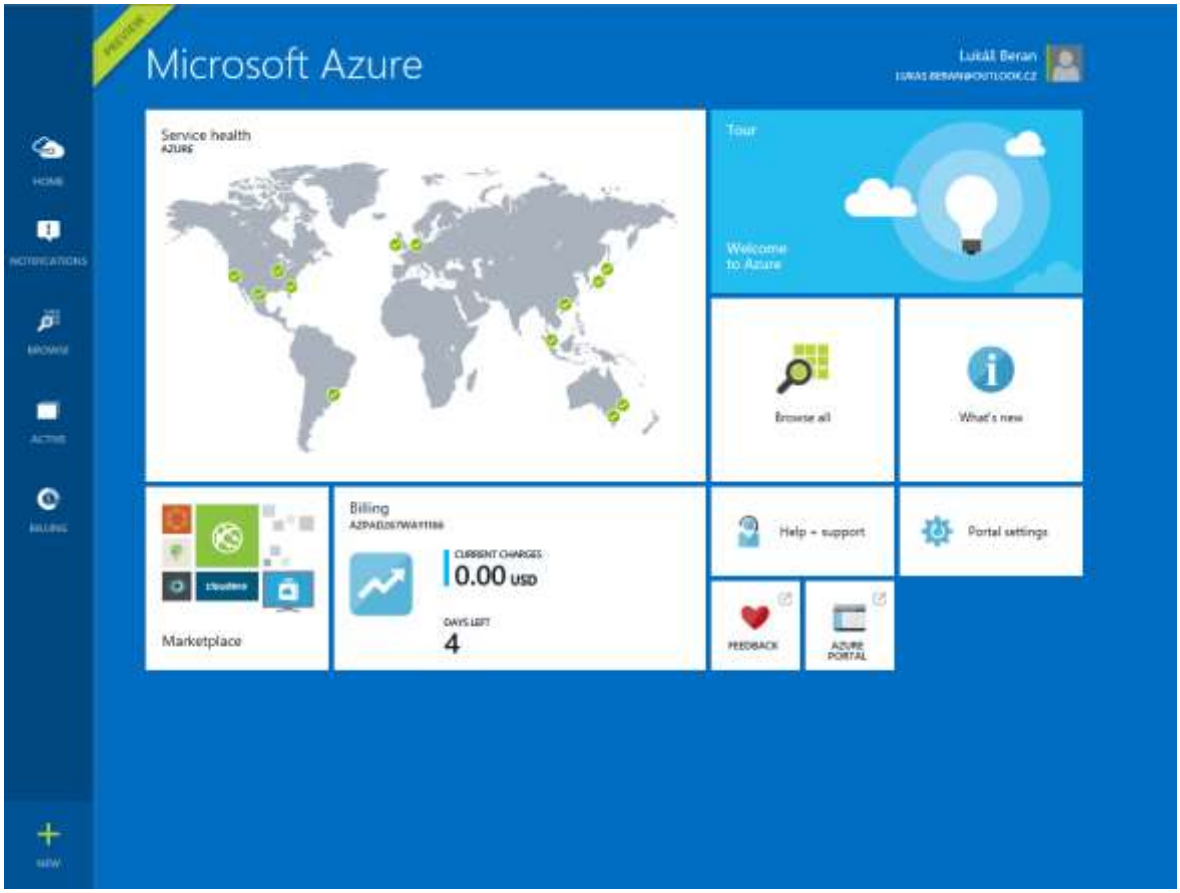
Portál Microsoft Azure, pod který spadá i Machine Learning, se nachází na adrese https://azure.microsoft.com/, která je domovskou stránkou pro všechny služby Microsoft Azure. Azure Machine Learning se pak nachází na adrese https://azure.microsoft.com/ml.

Pro první kontakt s Microsoft Azure a jeho Machine Learningem je vhodným zdrojem informací, kromě této diplomové práce, také oficiální dokumentace na adrese https://azure.microsoft.com/en-us/documentation/services/machine-learning/ a domovská stránka Azure ML Studia na adrese https://studio.azureml.net/, kde jsou kromě novinek uvedeny také příklady, příručka pro ML Studio a krátká videa s ukázkami. Posledním dobrým zdrojem informací může být Microsoft Virtual Academy na adrese http://www.microsoftvirtualacademy.com/, kde můžeme najít základní kurzy pro většinu produktů a služeb Microsoftu včetně Machine Learningu.

Vzhledem k tomu, že Machine Learning je stále ve fázi Preview, neexistuje k němu žádná oficiální podpora, na kterou by bylo možné zavolat nebo napsat. Pro podporu je však možné využít anglické komunitní fórum na adrese https://social.msdn.microsoft.com/forums/en-us/home?forum=MachineLearning, kde podporu poskytují jednak ostatní uživatelé, ale také přímo zaměstnanci společnosti Microsoft, nejedná se však o oficiální kanál podpory s garantovanou odezvou a řešením. Až bude Machine Learning finální, bude mu poskytována oficiální garantovaná podpora přes Azure Support.

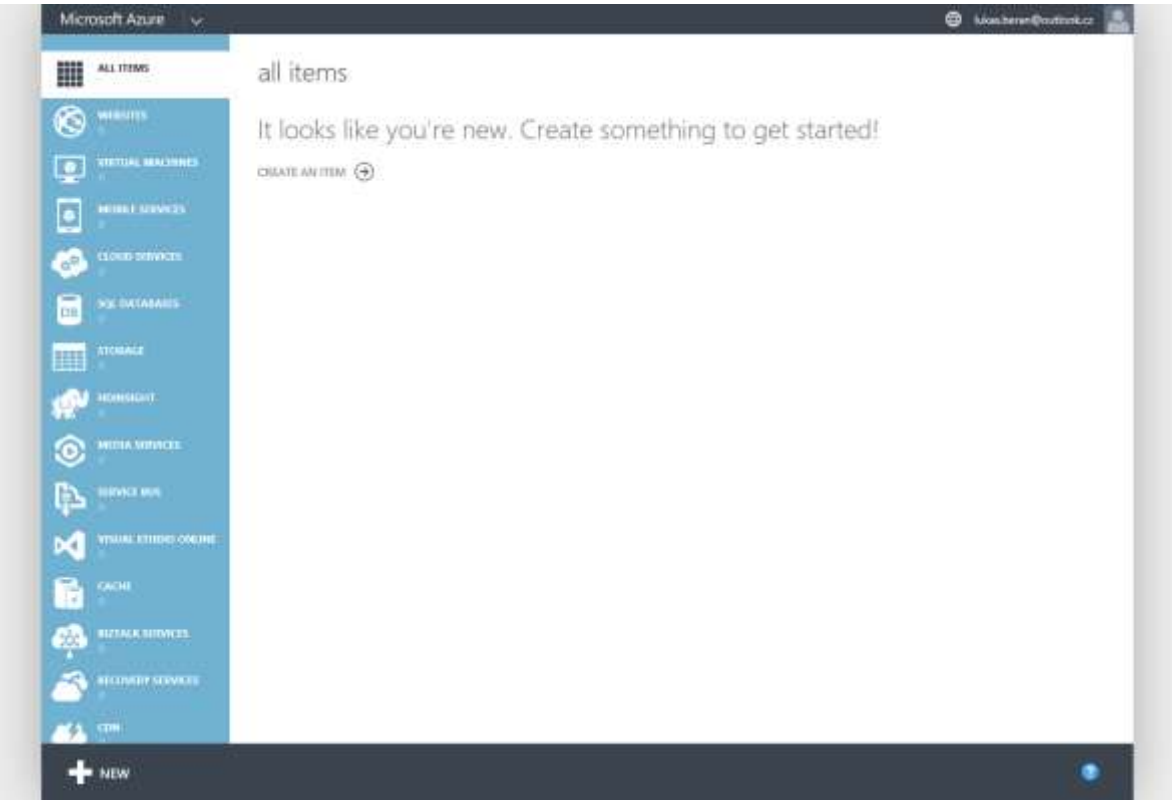
6.1 První kontakt s Azure Machine Learning

Do portálu je možné se přihlásit na adrese https://manage.windowsazure.com/. Pro přihlášení se používá Microsoft účet, kterému je možné pro vyzkoušení aktivovat testovací předplatné s přiřazeným kreditem 150 € (v době psaní diplomové práce) pro všechny Azure služby. Portál Microsoft Azure je v době psaní diplomové práce dostupný ve dvou verzích – základní starší verzi (Obrázek 4) a nové moderní verzi ve fázi Preview (Obrázek 3). V nové Preview verzi portálu v tuto chvíli ještě nejsou dostupné všechny služby, mimo jiné i Machine Learning, což by se ale mělo v blízké budoucnosti změnit a všechny služby by měly být postupně převedeny do nové verze portálu, která se poté stane hlavní a finální.



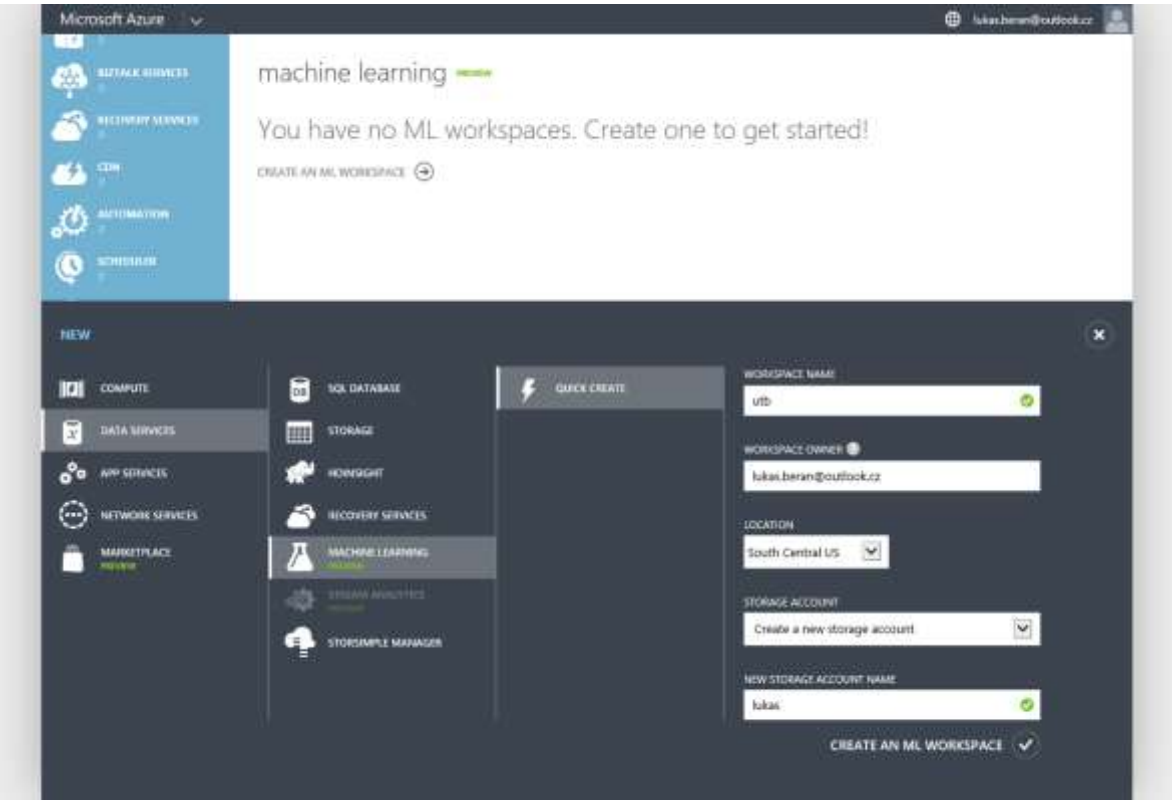
Obrázek 3 - Microsoft Azure Preview portál

Z tohoto důvodu je stále nutné používat starší verzi portálu, kde jsou dostupné všechny služby a tato verze je také výchozí.



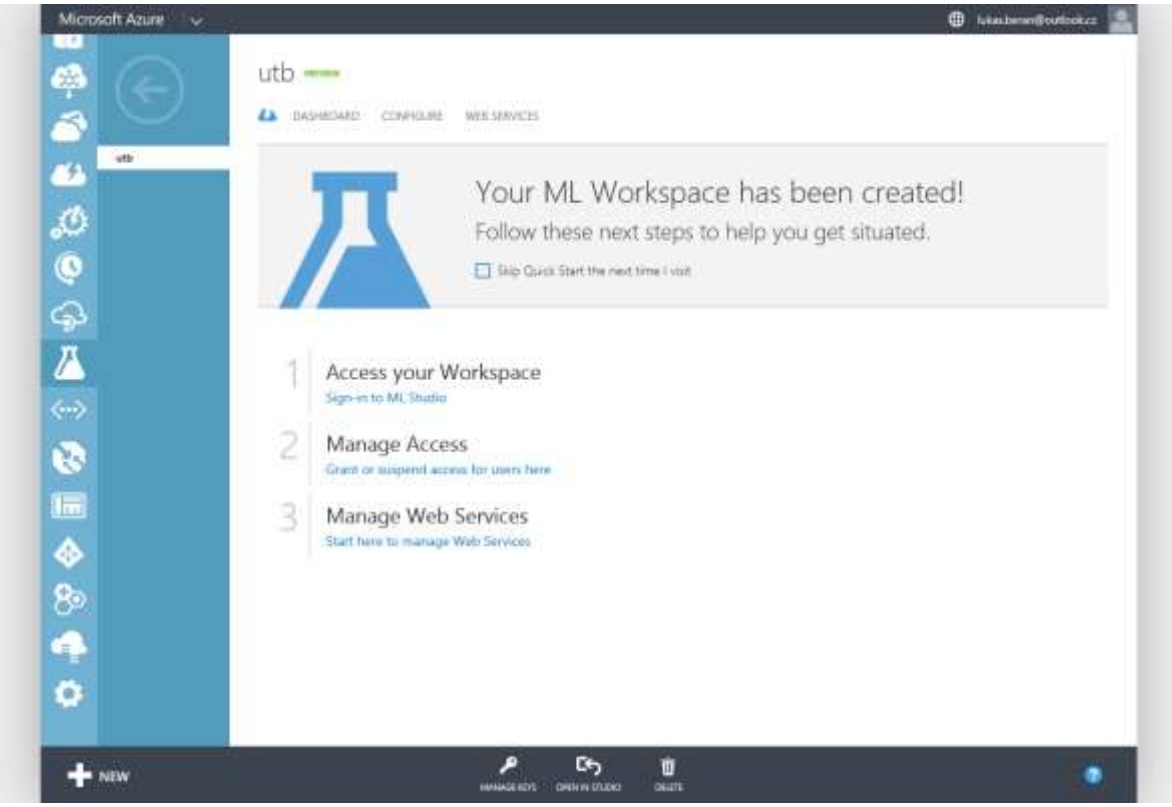
Obrázek 4 - Microsoft Azure portál

V levém menu se nachází nabídka všech služeb, které je možné vytvořit. Hlavní panel obsahuje seznam již vytvořených služeb. V levém menu tedy vybereme Machine Learning a vytvoříme novou pracovní plochu (workspace), jak je vidět na obrázku (Obrázek 5).



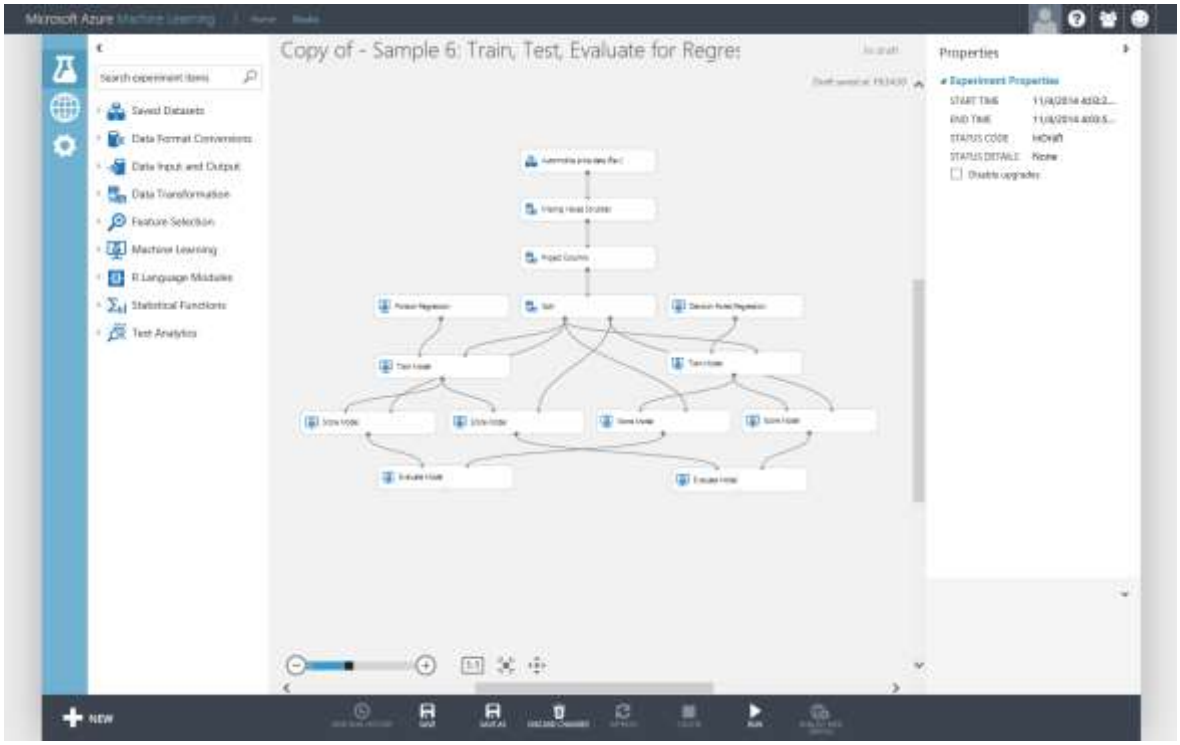
Obrázek 5 - Vytvoření nové pracovní plochy Machine Learning

Po vytvoření je možné vidět novou pracovní plochu v hlavní části okna. Kliknutím na název si danou pracovní plochu otevřeme a vidíme základní nabídku Machine Learningu (Obrázek 6) s možností spustit ML Studio, spravovat uživatele nebo spravovat webové služby Web Services. Ve spodní liště je možné spravovat klíče pro účty úložiště, otevřít ML Studio nebo odstranit otevřenou pracovní plochu.



Obrázek 6 - Workspace v Microsoft Azure

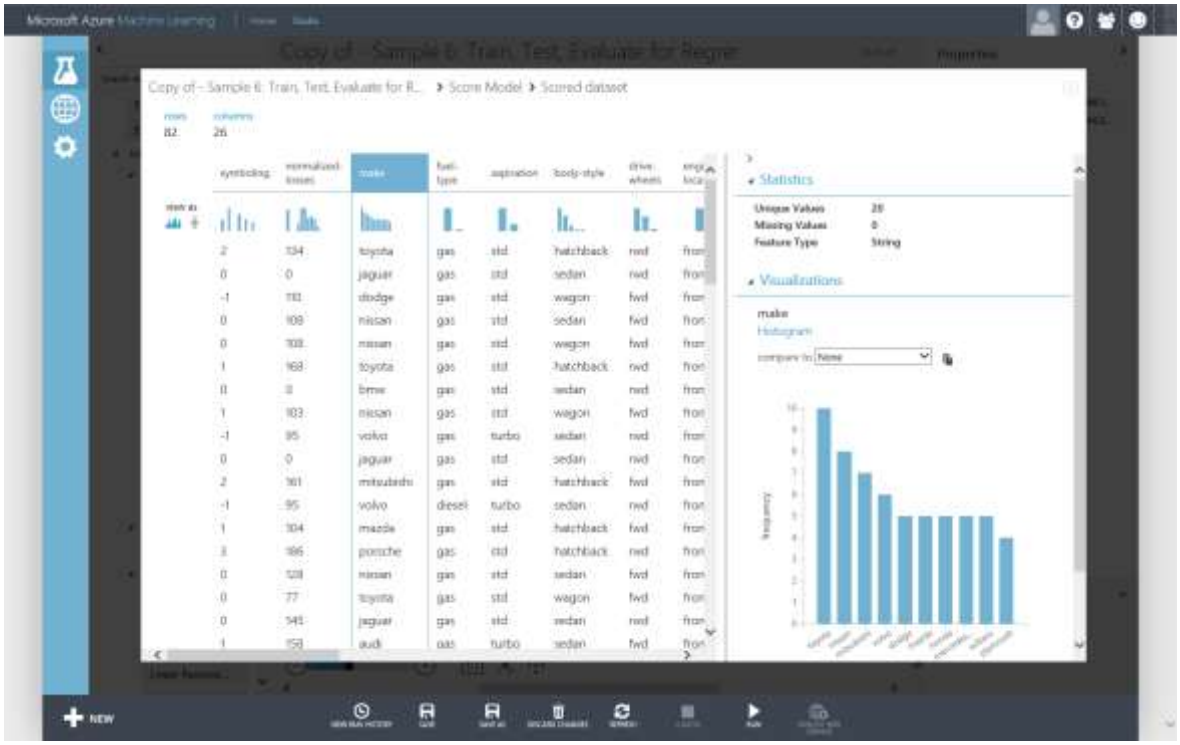
Jako první je vhodné si otevřít ML Studio, jelikož od toho se odvíjí i všechny další možnosti. Při prvním spuštění ML Studia se spustí krátké uvítací video se shrnutím a ukázkou možností použití. K dispozici je hned několik ukázkových experimentů (Obrázek 7) a datasetů, na kterých je možné si vyzkoušet použití ML Studia.



Obrázek 7 - Testovací experiment v ML Studiu

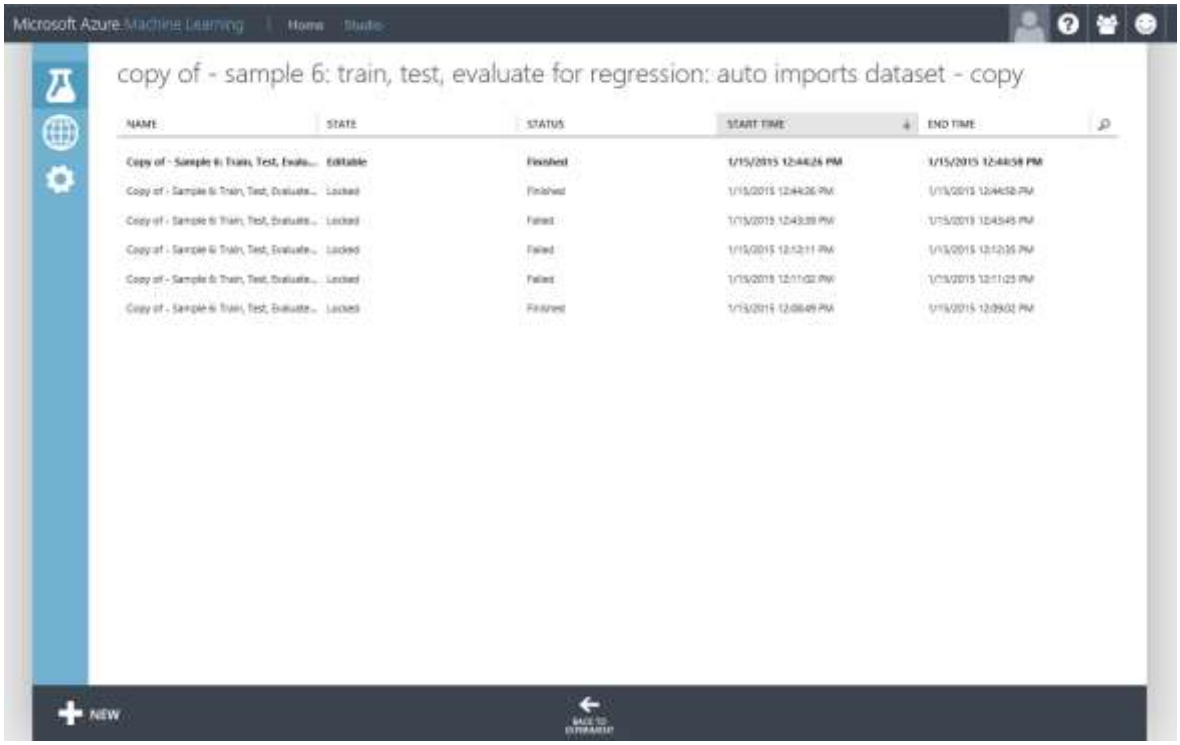
Po spuštění experimentu je možné vidět v reálném čase zpracovávání jednotlivých částí modelu. Model je možné spravovat přímo pomocí drag&drop. V levé nabídce se nachází všechny moduly, které je možné do modelu přidat. Ke všem modulům modelu je k dispozici nápověda, ve které je popis, co daný modul dělá, dále pak očekávané vstupy, výstupy a možné výjimky s jednotlivými chybovými kódy, které mohou nastat.

Ke každému modulu je možné přidávat komentáře nebo se podívat do logu. Další nabídky jsou obsaženy přímo na jednotlivých spojeních mezi moduly a na výstupních bodech všech modulů, kde je k dispozici vizualizace (Obrázek 8) a další práce s daty.



Obrázek 8 - Vizualizace dat v ML Studiu

U každého experimentu je možné vidět historii jeho běhů. V této historii je vidět i konfigurace daného experimentu a jeho stav, jakým skončil (Obrázek 9).



Obrázek 9 - Historie běhu experimentu

6.2 Vytvoření vlastního experimentu

Experimenty se skládají z částí pro tvorbu modelu, trénování modelu, ohodnocení a testování modelu. Kombinací tedy můžeme vytvořit experiment, který vezme data, vytrénuje na nich model a tento model poté aplikuje na nová data. Pokud jsou vstupní data nějak nevhodně uspořádána, je možné použít modul pro předzpracování těchto dat, aby byla pro předpovědi vhodnější a lépe zpracovatelná. Možné je například některá data oddělit pro trénování a jiná pro ověření modelu nebo nepotřebná data úplně odstranit.

Vytvoření experimentu v ML Studiu se může skládat ze základních pěti kroků rozdělených do tří částí:

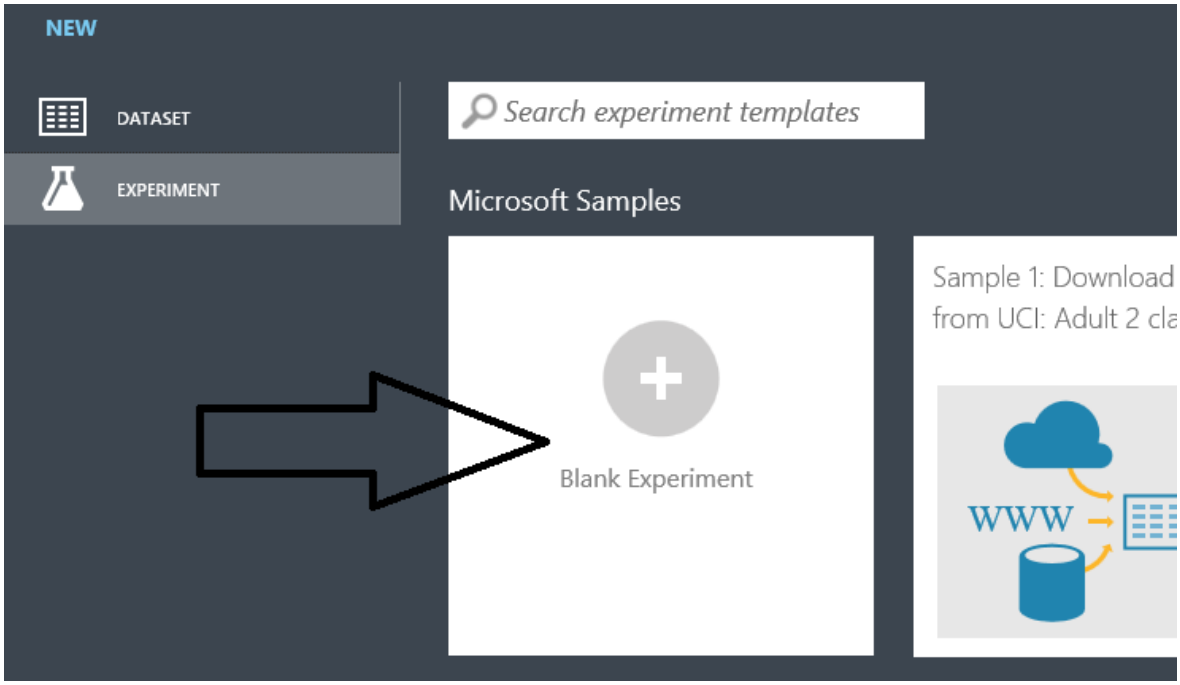
- Tvorba modelu
    1. Získání dat
    2. Předzpracování dat
    3. Definování parametrů
- Trénování modelu
    4. Výběr a aplikace učícího algoritmu
- Ohodnocení a testování modelu
    5. Předpovědi nad novými daty

V následujícím praktickém příkladu se nachází ukázka vytvoření regresního modelu s použitím vzorových dat o automobilech. Cílem bude předpovědět cenu automobilu z jeho různých parametrů, například výrobce nebo technická specifikace automobilu.

6.2.1 Získání dat

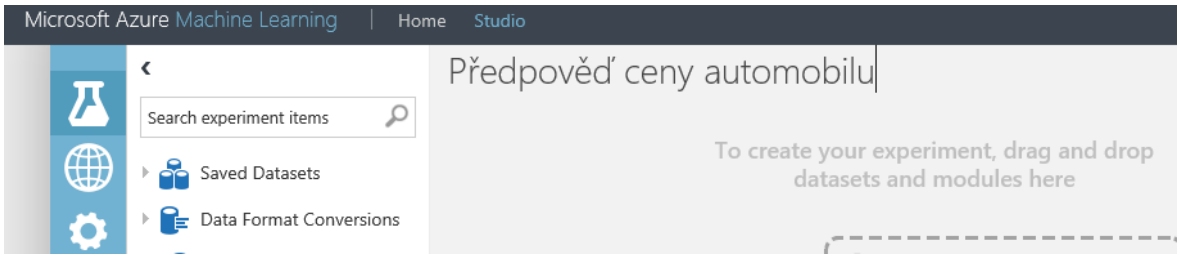
Vzorová data jsou již předpřipravena jako příklad v ML Studiu. Možné je však snadno nahrát i vlastní data například z CSV souboru.

V ML Studiu na spodní liště zvolíme NEW a vybereme EXPERIMENT – Blank Experiment (Obrázek 10).



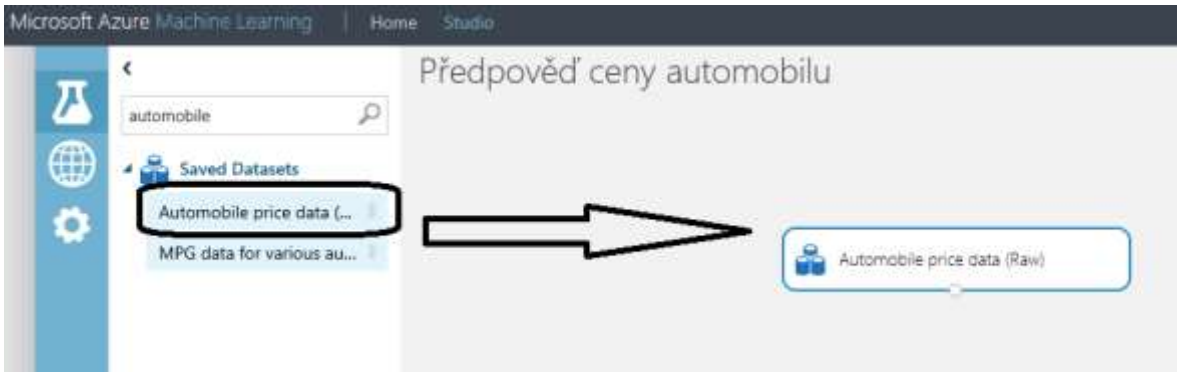
Obrázek 10 - Tvorba nového experimentu

Nový experiment si pojmenujeme, například *Předpověď ceny automobilu* (Obrázek 11).



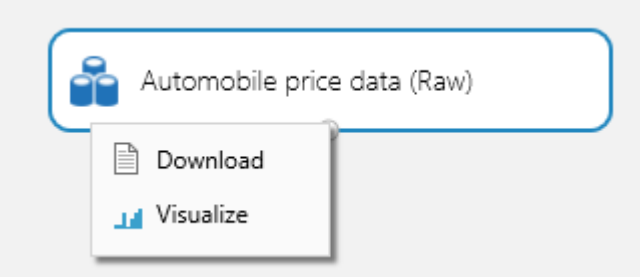
Obrázek 11 - Pojmenování experimentu

V levém sloupci se nachází nabídka s položkami experimentů, která obsahuje i již předpřipravené příklady dat. Vyhledáme tedy *automobile* a přetáhneme ho do pracovní plochy experimentu (Obrázek 12).



Obrázek 12 - Vybrání dat pro experiment

Nyní je možné se podívat, co je obsahem daného datasetu. Kliknutím pravým tlačítkem myši na kolečko u datasetu se otevře nabídka, kde je možné data stáhnout nebo vizualizovat (Obrázek 13).



Obrázek 13 - Volba u dat

Výběrem možnosti vizualizace se hodnoty z datasetu zobrazí v tabulce, kde každá instance automobilu je na jednom řádku, a v sloupcích jsou parametry každého automobilu (Obrázek 14). Jedním z parametrů je i cena, kterou se budeme snažit předpovídat.

Předpověď ceny automobilu › Automobile price data (Raw) › dataset

rows 205 columns 26

view as

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	length	width	hei
	3		alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.
	3		alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.
	1		alfa-romero	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.
	2	164	audi	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.
	2	164	audi	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.
	2		audi	gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.
	1	158	audi	gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.
	1		audi	gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.
	1	158	audi	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.
	0		audi	gas	turbo	two	hatchback	4wd	front	99.5	178.2	67.9	52
	2	192	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.
	0	192	bmw	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.
	0	188	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.
	0	188	bmw	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.
	1		bmw	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.
	0		bmw	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.
	0		bmw	gas	std	two	sedan	rwd	front	103.5	193.8	67.9	53.
	0		bmw	gas	std	four	sedan	rwd	front	110	197	70.9	56

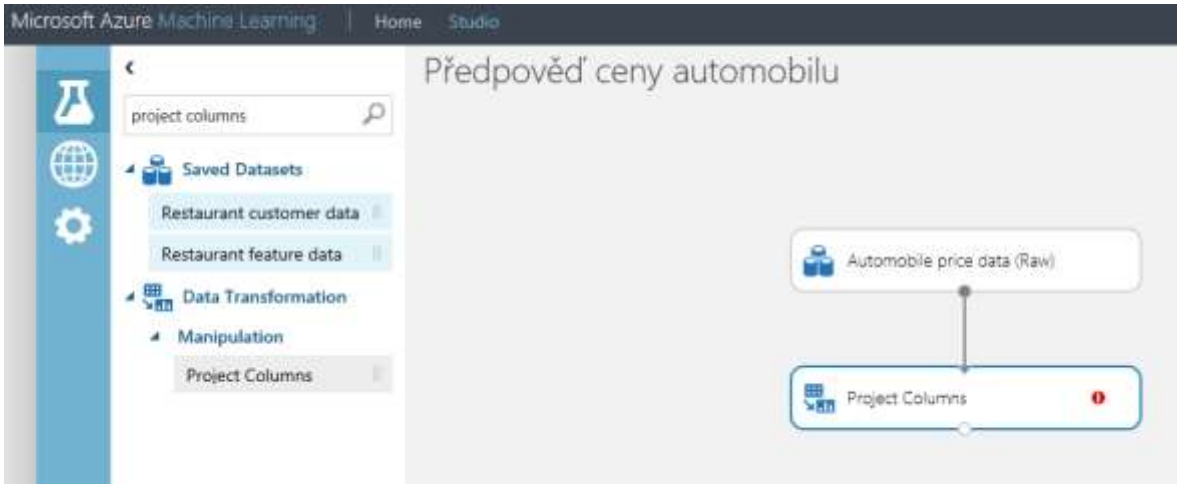
Obrázek 14 - Vizualizace dat z datasetu

6.2.2 Předzpracování dat

Datasety obvykle vyžadují určitou formu předzpracování před tím, než mohou být analyzovány. Některé hodnoty například nemusí být zadané, přitom pro analýzu potřebujeme mít data kompletní. Proto v našem případě odstraníme všechny řádky, které mají nějaké nezadané hodnoty parametrů. Sloupec *normalized-losses* má velmi mnoho nezadaných hodnot, proto je vhodné ho odstranit celý, aby nedošlo k odstranění velkého počtu řádků jen kvůli jednomu parametru.

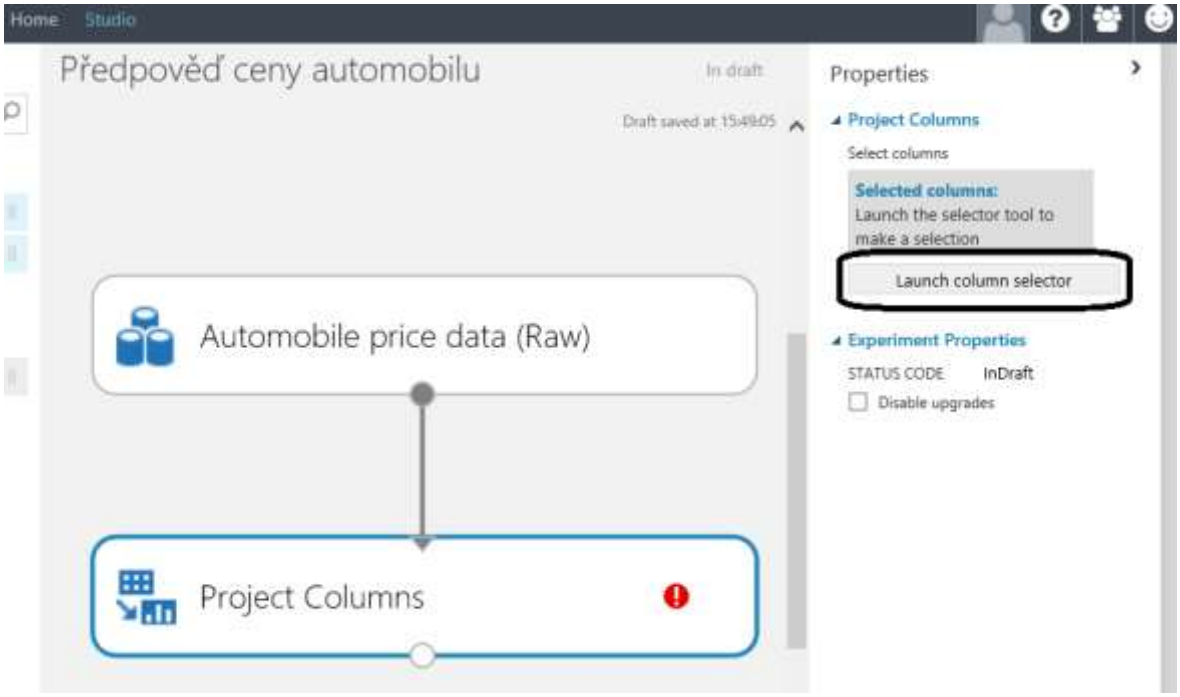
Odstranění chybějících hodnot ze vstupních dat je předpokladem pro použití většiny modulů ML Studia.

Jako první tedy odstraníme sloupec *normalized-losses*. V levé nabídce si vyhledáme modul *Project Columns*, který umožní odstranit nežádoucí sloupce, a přetáhneme ho do pracovní plochy experimentu pod dataset automobilů, k jehož výstupu modul připojíme (Obrázek 15).



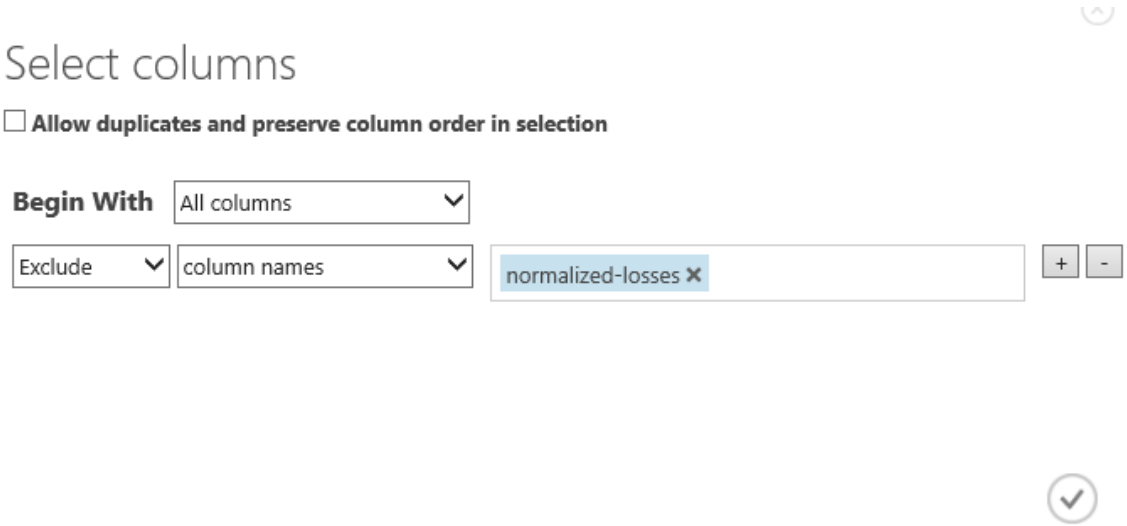
Obrázek 15 - Připojení modulu pro odstranění sloupců

U modulu je vidět varování, že je vyžadováno zadání hodnoty. Označíme tedy modul a v pravé nabídce zvolíme *Launch column selector* (Obrázek 16).



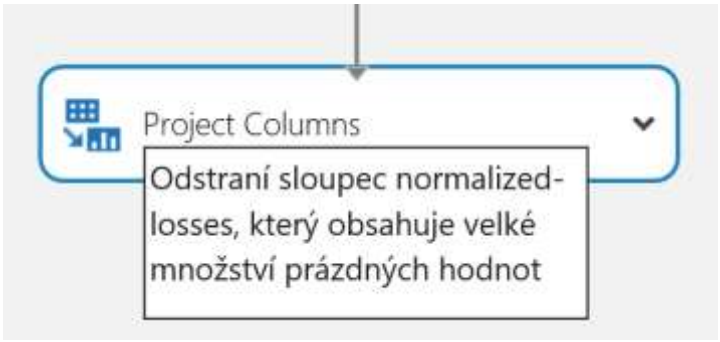
Obrázek 16 - Výběr sloupců pro odstranění

Ujistíme se, že v nabídce *Begin with* je vybrána možnost *All columns*, což zajistí, že všechny sloupce kromě níže zvolených budou ponechány v tabulce. Sloupce, které chceme odstranit, zadáme vybráním možnosti *Exclude* a *Column Names* v roletkách níže a následně vypsáním názvů sloupců pro odstranění (Obrázek 17).



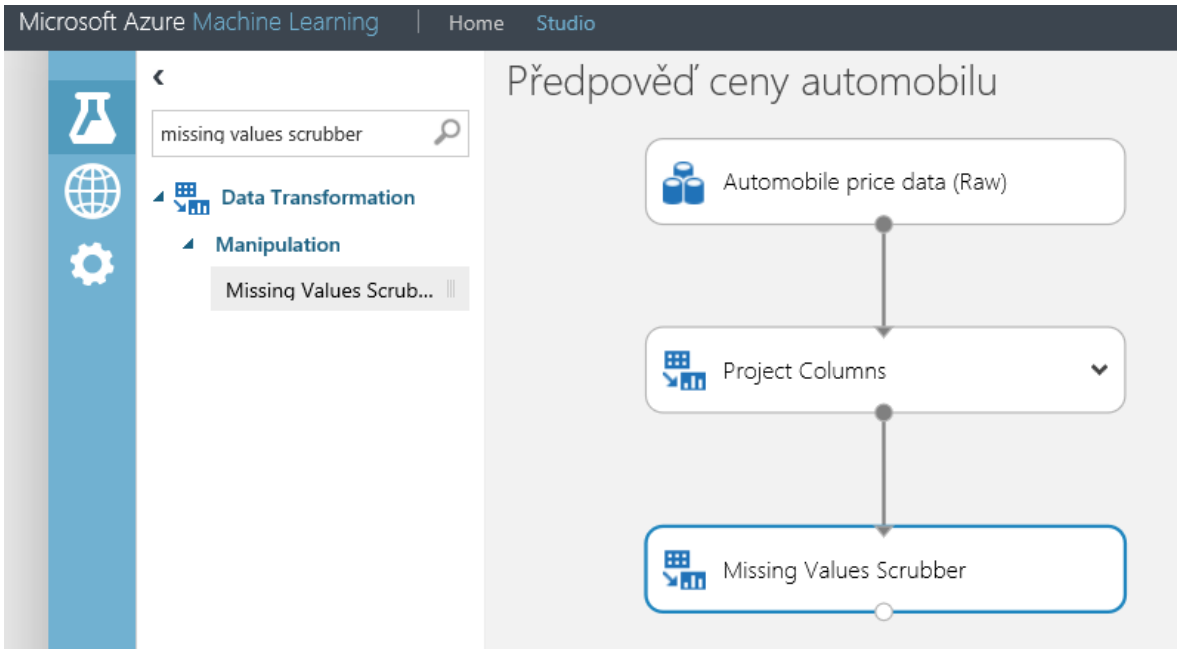
Obrázek 17 - Výběr sloupců pro odstranění

U modulů je možné přidávat komentáře. Pomocí dvojkliku na modul se otevře okno, kde je možné zadat komentář pro bližší popis, co daný modul s daty opravdu dělá. V našem případě si můžeme přidat například komentář „Odstraní sloupec normalized-losses, který obsahuje velké množství prázdných hodnot“ (Obrázek 18).



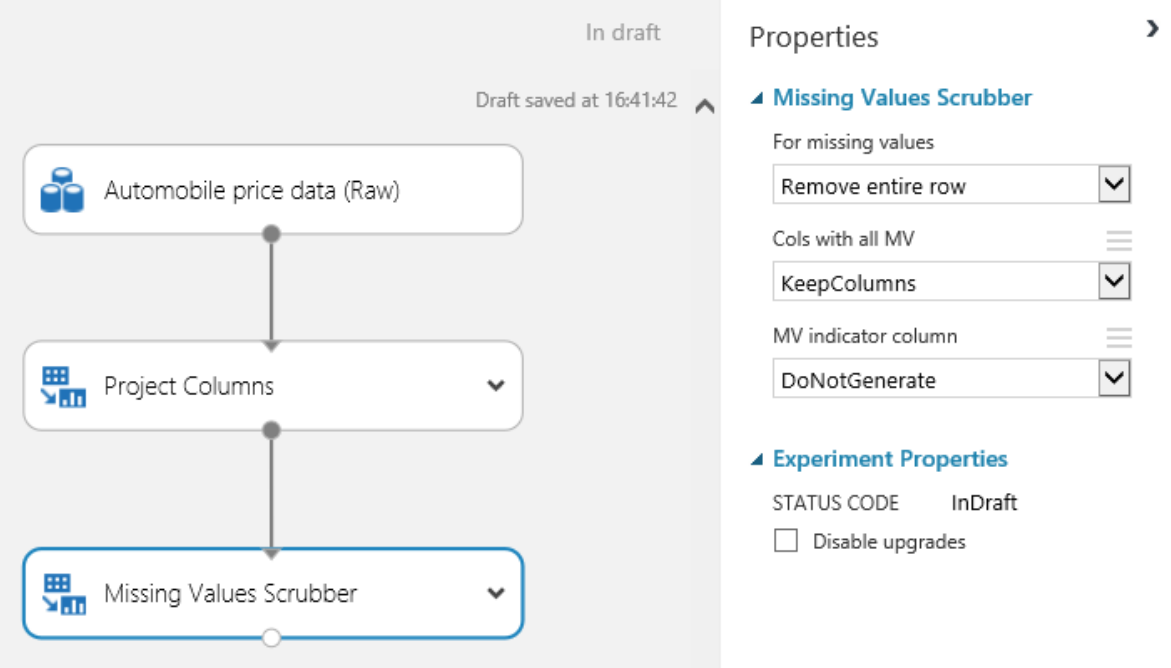
Obrázek 18 - Přidání komentáře k modulu

Nyní přidáme další modul, který odstraní řádky obsahující nějaké prázdné hodnoty. Tento modul se jmenuje *Missing Values Scrubber*, najdeme ho opět v levé nabídce a přidáme pod *Project Columns* (Obrázek 19).



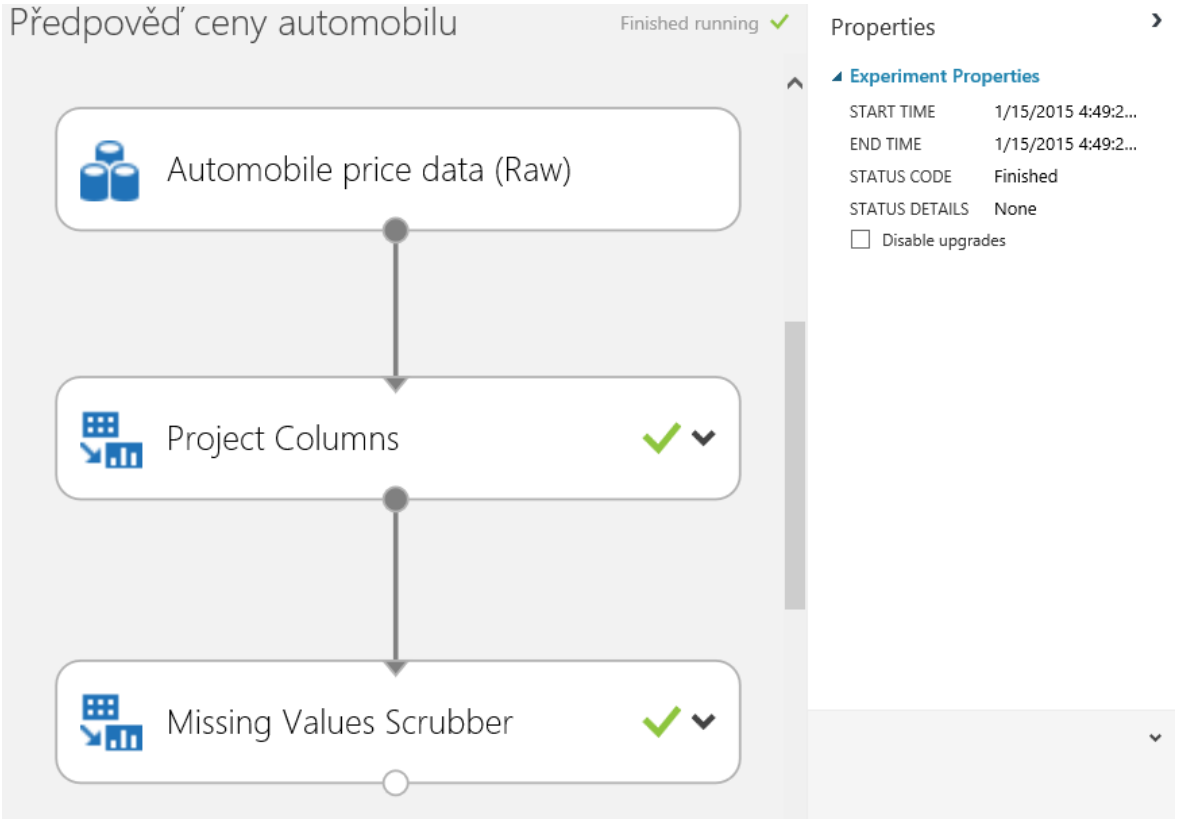
Obrázek 19 - Přidání modulu pro odstranění řádků s prázdnými hodnotami

Po označení modulu v pravé nabídce vybereme odstranění celého řádku u chybějících hodnot (Obrázek 20) a můžeme opět přidat komentář.



Obrázek 20 - Odstranění řádků s prázdnými hodnotami

Tímto jsme připraveni vyčistit data od nežádoucích položek. Experiment spustíme tlačítkem *Run* na spodní liště. U jednotlivých modulů je nejdřív vidět symbol hodin, který značí, že daná operace je ve frontě a čeká. Při běhu operace je u daného modulu zobrazen symbol točícího se kolečka a po úspěšném dokončení je vidět zelená „fajfka“. V pravé nabídce jsou vidět informace o běhu experimentu (Obrázek 21).



Obrázek 21 - Dokončený experiment

Upravenou tabulku s již odstraněnými hodnotami je možné zobrazit kliknutím pravým tlačítkem myši na spodní tečku modulu *Missing Values Scrubber*, kde je opět možnost vizualizace nebo stažení dat. Vidíme, že z ukázkového datasetu zmizelo 12 řádků, které obsahovaly nějaké prázdné hodnoty, a jeden sloupec *normalized-losses* (Obrázek 22).

Předpověď ceny automobilu › Missing Values Scrubber › Results dataset

rows 193 columns 25

	symboling	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	length	width	height	curb wei
view as													
	3	alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	254
	3	alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	254
	1	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4	282
	2	audi	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3	233
	2	audi	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3	282
	2	audi	gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.1	250
	1	audi	gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.7	284
	1	audi	gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.7	295
	1	audi	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.9	308
	2	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	239
	0	bmw	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	239
	0	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	279
	0	bmw	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	276
	1	bmw	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	305
	0	bmw	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	323
	0	bmw	gas	std	two	sedan	rwd	front	103.5	193.8	67.9	53.7	338
	0	bmw	gas	std	four	sedan	rwd	front	110	197	70.9	56.3	350
	2	chevrolet	gas	std	two	hatchback	fwd	front	88.4	141.1	60.3	53.2	148

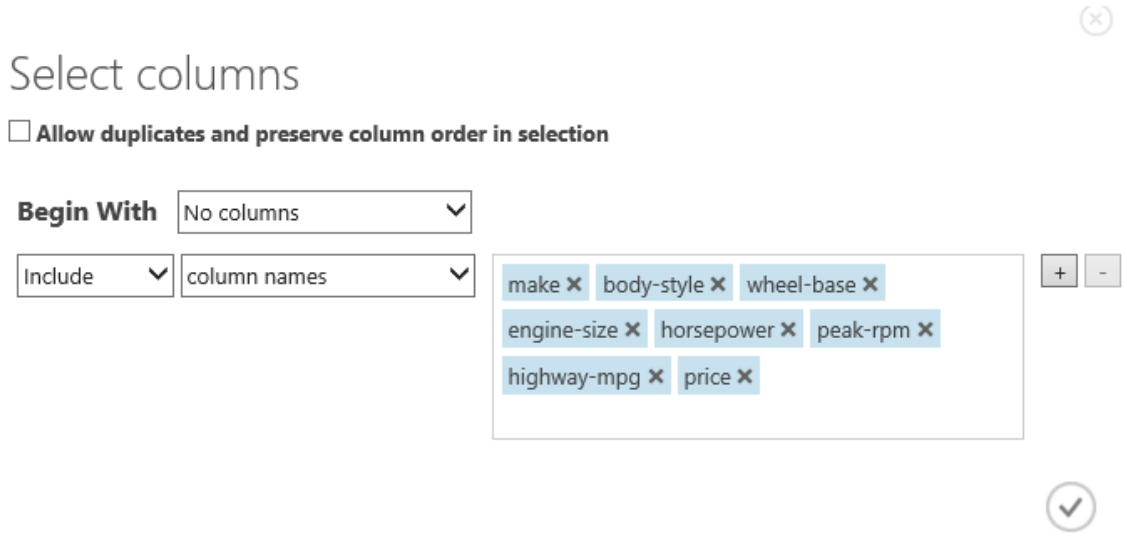
Obrázek 22 - Předpřipravený dataset

6.2.3 Definování parametrů

Ve strojovém učení jsou parametry individuální měřitelné vlastnosti něčeho, co nás zajímá. V našem datasetu každý řádek reprezentuje jeden automobil a sloupce reprezentují parametry tohoto automobilu. Nalezení vhodné skupiny parametrů pro tvorbu prediktivního modelu vyžaduje experimentování a znalost daného problému. Některé parametry jsou totiž vhodnější pro tvorbu předpovědi než jiné. Stejně tak některé parametry mají silnější korelaci s jinými parametry. V našem případě je to například spotřeba ve městě (*city-mpg*) se spotřebou na dálnici (*highway-mpg*), proto tyto parametry společně neposkytnou mnoho informací našemu modelu a jeden z nich můžeme vynechat.

Nyní je možné vytvořit model s využitím části parametrů upraveného datasetu. V tuto chvíli již začíná experimentování, takže je možné, že na první pokus nemusí vyjít správné výsledky. V takovém případě je vhodné upravit parametry modelu a zkusit spustit experiment znovu.

Jako první zkusíme zvolit parametry *make* (značka automobilu), *body-style* (typ karoserie), *wheel-base* (vzdálenost mezi přední a zadní nápravou), *engine-size* (objem motoru v kubických palcích – CID), *horsepower* (výkon motoru v koňských silách – HP), *peak-rpm* (maximální otáčky motoru – RPM), *highway-mpg* (spotřeba na dálnici v mílích na galon – MPG) a nakonec *price* (cena automobilu v amerických dolarech – USD). Cenu potřebujeme znát pro trénování našeho modelu. Tyto parametry zadáme do modulu *Project Columns*, který připojíme za modul *Missing Values Scrubber*. Modulu můžeme opět přidat komentář. Modul si označíme a v pravé nabídce klikneme na *Launch column selector*, kde zvolíme *No columns* u možnosti *Begin with*. Dále zvolíme *Include* a *Column Names* a zadáme jména sloupců (parametrů), které chceme zahrnout (Obrázek 23). Tím jsme docílili, že budou zpracovány pouze vybrané sloupce a ostatní budou ignorovány. Protože jsme již spustili daný experiment a tento modul je připojený až za *Missing Values Scrubber*, máme zobrazený již vyčištěný dataset.



Obrázek 23 - Výběr parametrů pro předpověď

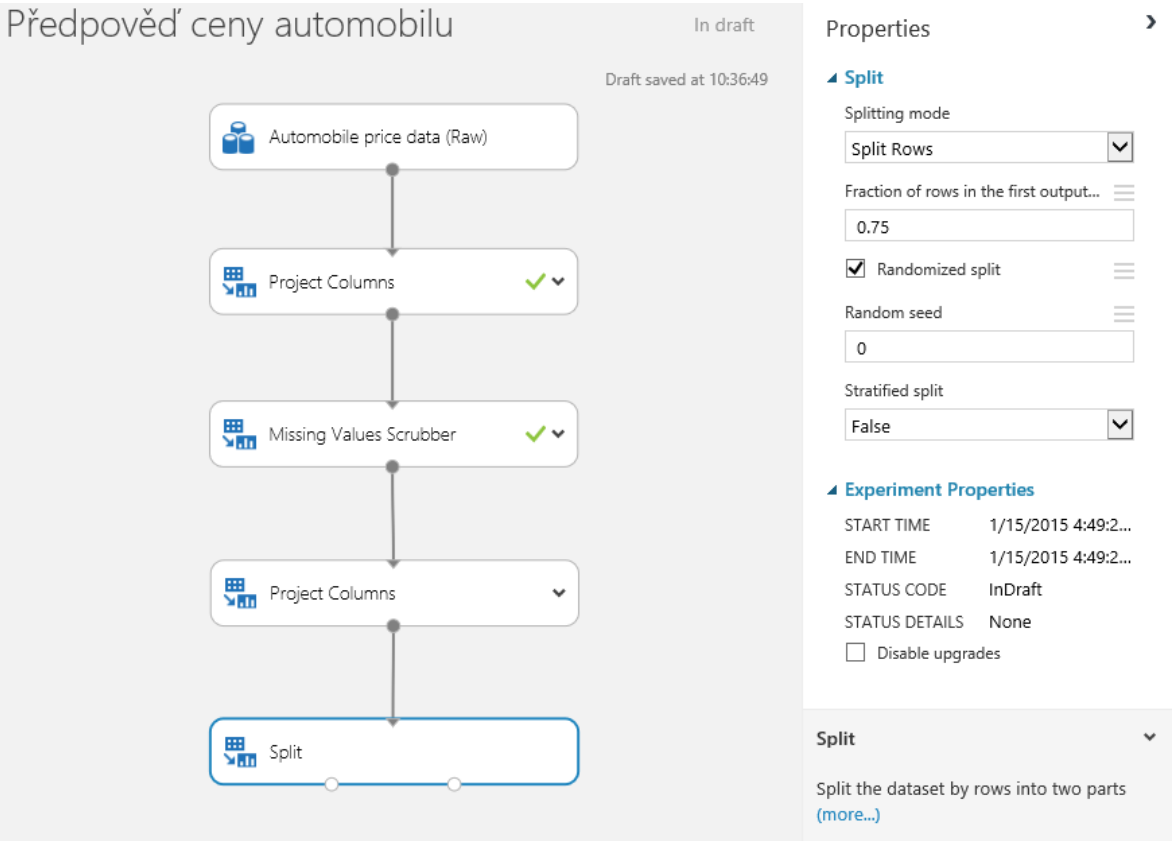
Tímto dostaneme upravený dataset, který následně použijeme pro trénování modelu.

6.2.4 Výběr a aplikace učícího algoritmu

Nyní jsou data připravená pro další zpracování, kterým je trénování a otestování prediktivního modelu. Základními technikami strojového učení jsou klasifikace a regrese. Zjednodušeně lze říci, že klasifikace se používá pro tvorbu předpovědí ze zvoleného souboru hodnot, jako například barva (výběr z několika konkrétních hodnot), a regrese se používá pro tvorbu předpovědí z kontinuálního souboru hodnot, jako například věk člověka (libovolné hodnoty).

My chceme předpovídat cenu automobilu, což může být libovolná hodnota, proto použijeme regresní model. V tomto příkladu budeme trénovat jednoduchý lineární regresní model a v dalším kroku otestujeme jeho správnost. Pokud bychom nebyli s přesností spokojeni, můžeme použít jiný typ modelu a test opakovat.

1. **Rozdělení dat na trénovací a testovací.** K tomu použijeme modul *Split*, který připojíme k poslednímu modulu *Project Columns*. U tohoto modulu nastavíme pouze poměr, jakým má rozdělit data na naše trénovací a testovací. Zvolíme 0,75, abychom měli 75 % dat pro trénování a zbylých 25 % pro testování (Obrázek 24).
   Změnou hodnoty *Random seed* můžeme vytvořit jiné náhodné příklady pro trénování a testování. Více informací z [11].

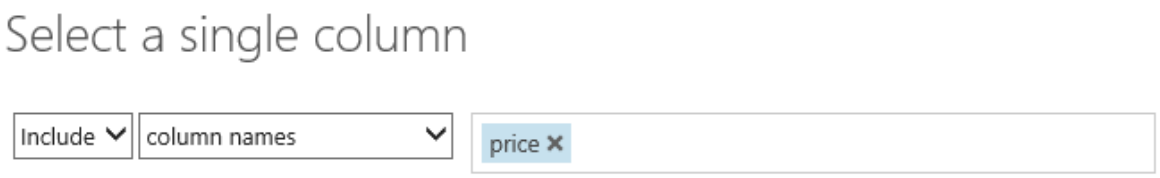


Obrázek 24 - Rozdělení hodnot na trénovací a testovací

2. **Spuštění experimentu.** Tímto dojde k aplikaci nově přidaných modulů.
3. **Výběr učícího algoritmu.** Pro výběr vhodného učícího algoritmu si v levé nabídce otevřeme kategorii *Machine Learning – Initialize Model*. Zde je vidět několik dalších kategorií modulů, které mohou být použity pro inicializaci učícího algoritmu. V tomto příkladu vybereme modul *Linear Regression* v kategorii *Regression*. Modul umístíme vedle modulu *Split*, ale nebudeme ho zatím připojovat (Obrázek 25).

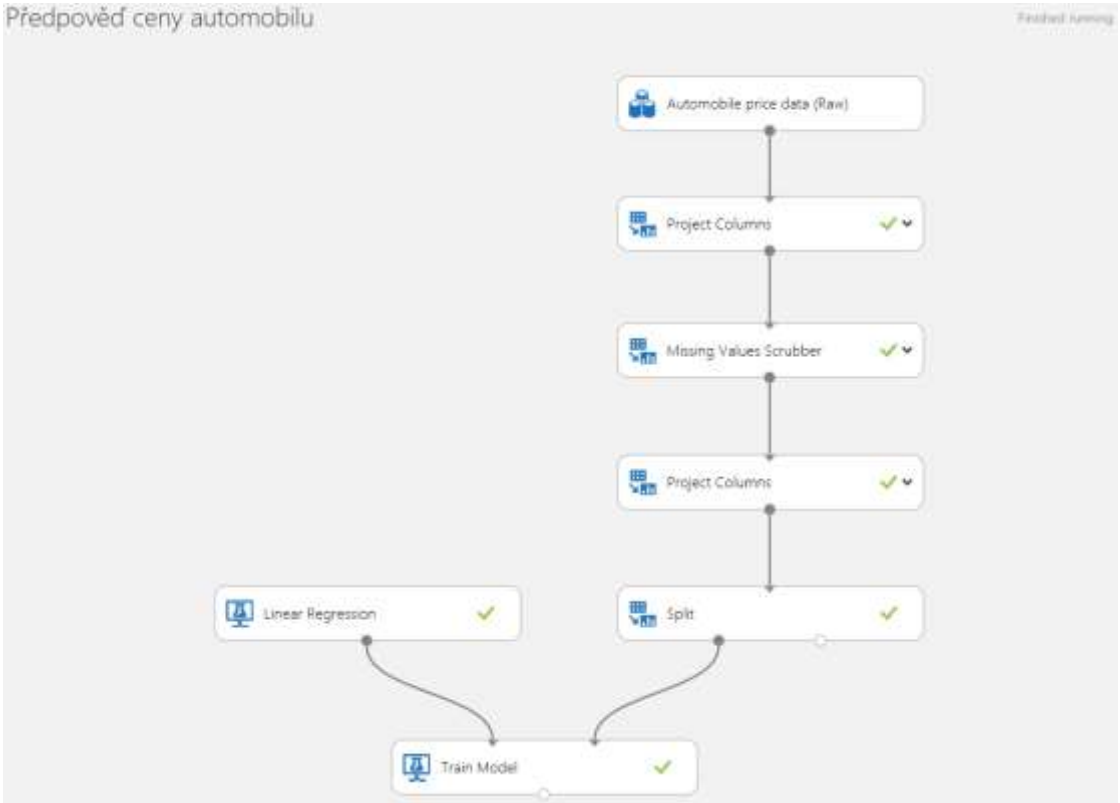
Obrázek 25 - Přidání modulu lineární regrese

4. **Modul pro trénování.** V levé nabídce modulů si najdeme modul pro trénování modelu *Train Model*. Modul připojíme vlevo na modul lineární regrese a vpravo na modul rozdělení dat. V modulu pro trénování modelu je potřeba vybrat data, která má daný model předpovídat. V pravé nabídce zvolíme *Launch column selector* a následně *Include*, *Column Names* a vybereme sloupec *price* (Obrázek 26).



Obrázek 26 - Výběr dat, které má model předpovídat

5. **Spuštění experimentu.** Výsledkem je natrénovaný regresní model (Obrázek 27), který může být použit pro ohodnocení nových příkladů pro tvorbu predikcí.

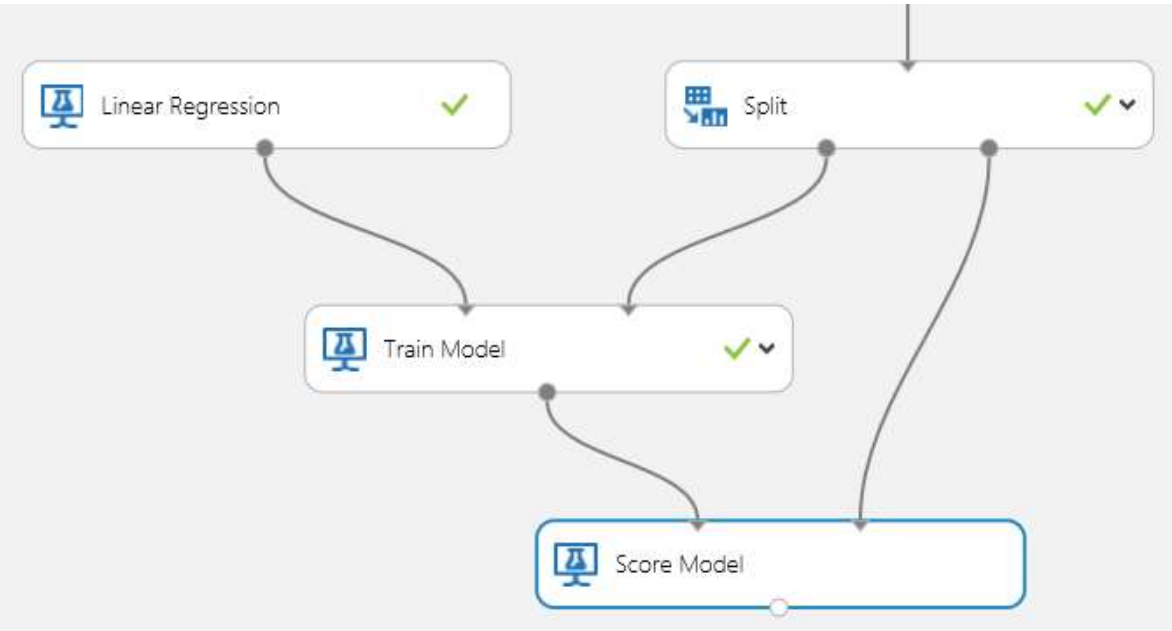


Obrázek 27 - Natrénovaný regresní model pro tvorbu predikcí

6.2.5 Předpovědi nad novými daty

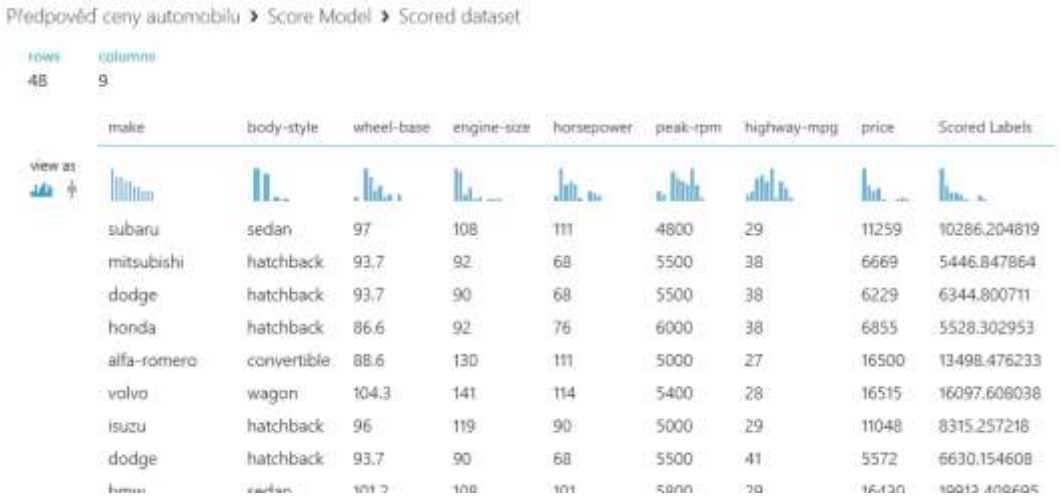
Nyní máme natrénovaný model a můžeme přistoupit k ohodnocení zbylých 25 % dat, abychom otestovali, jak dobře náš model dokáže předpovídat.

1. **Ohodnocení modelu.** V levé nabídce najdeme modul *Score Model* a přidáme ho pod *Train Model*. Vlevo ho připojíme k modulu *Train Model* a vpravo k modulu *Split*, kde máme zbylých 25 % dat (Obrázek 28).



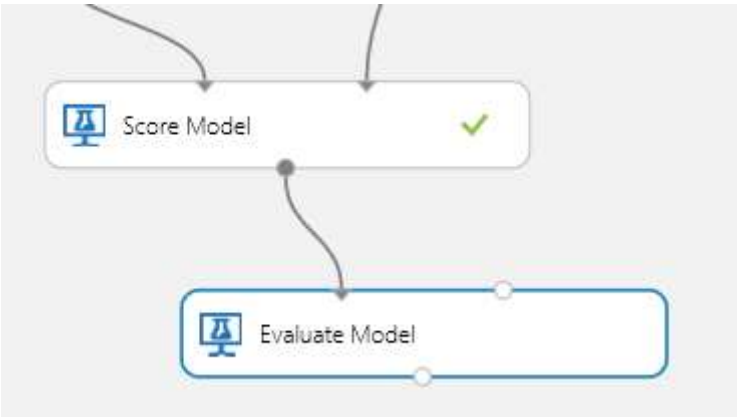
Obrázek 28 - Ohodnocení modelu

2. **Spuštění experimentu.** Spuštěním experimentu dojde k ohodnocení testovacích dat. Výsledek je možné vidět na výstupu modulu *Score Model* pomocí vizualizace nebo stažení dat, kde je zobrazena předpovězená cena automobilu (sloupec *Scored Labels*) spolu se skutečnou cenou (Obrázek 29).



Obrázek 29 - Výsledek ohodnocení modelu

3. **Vyhodnocení modelu.** Součástí každé předpovědi by mělo být vyhodnocení, jak se předpovězené hodnoty liší od skutečných. K tomu slouží modul *Evaluate Model*, který přidáme pod modul *Score Model* a levým vstupem ho připojíme do výstupu modulu *Train Model* (Obrázek 30). *Evaluate Model* má dva vstupy, protože umožňuje porovnat dva různé modely.

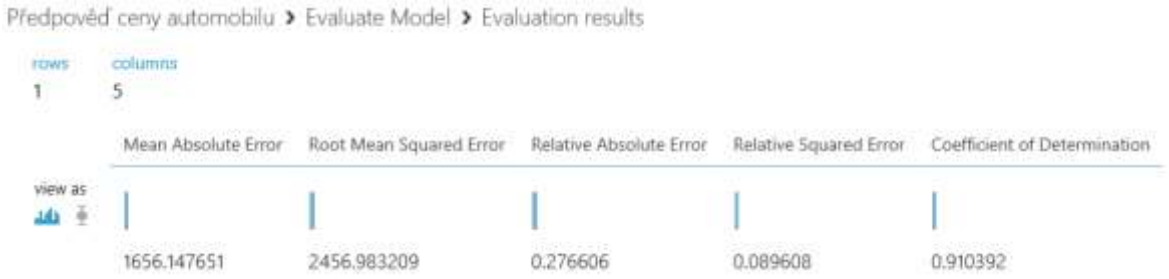


Obrázek 30 - Vyhodnocení modelu

4. **Spuštění experimentu.** Nyní spustíme experiment, abychom mohli vyhodnotit jeho výsledky pomocí nově přidaného modulu *Evaluate Model*, kde výsledky můžeme vizualizovat nebo stáhnout (Obrázek 31).
    - **Mean Absolute Error** – Průměr absolutních rozdílů mezi předpovězenou hodnotou a skutečnou hodnotou.
    - **Root Mean Squared Error** – Odmocnina průměru kvadratických odchylek předpovědi provedené na testovacím datasetu.

- **Relative Absolute Error** – Průměr absolutních odchylek vzhledem k absolutnímu rozdílu mezi skutečnými hodnotami a průměrem všech skutečných hodnot.
- **Relative Squared Error** – Průměr kvadratických odchylek vzhledem ke kvadratickému rozdílu mezi skutečnými hodnotami a průměrem všech skutečných hodnot.
- **Coefficient of Determination** – Známý také jako „R na druhou“. Jedná se o statistický údaj ukazující vhodnost modelu k datům.

Pro všechny statistické údaje platí, že menší je lepší, tedy že předpovězené hodnoty lépe odpovídají skutečným hodnotám. Naopak u koeficientu determinace platí, že čím více se blíží k hodnotě jedna (1), tím lepší je předpověď.



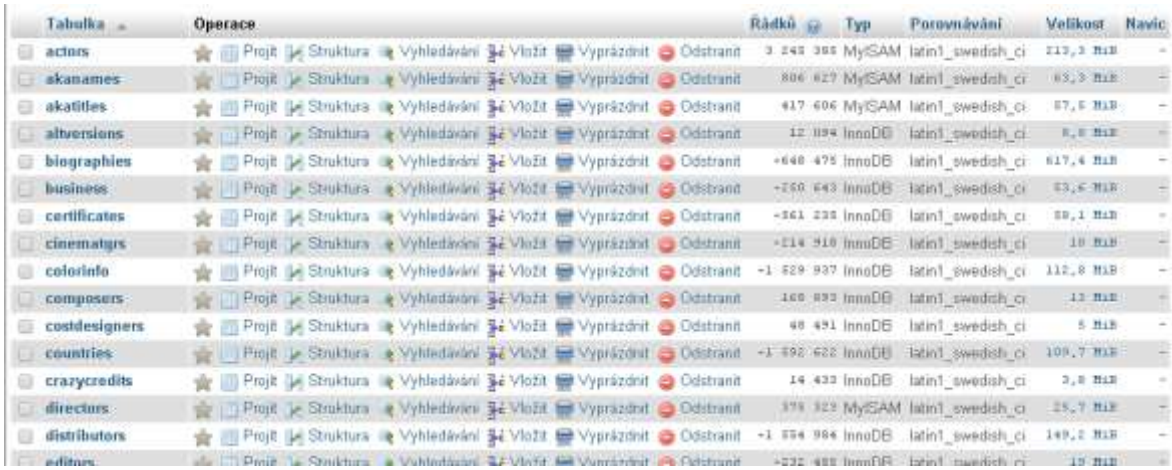
Obrázek 31 - Statistické informace o modelu

Jak vidíme ze statistických informací o modelu na obrázku (Obrázek 31) a tabulky výsledků z obrázku (Obrázek 29), výsledky nejsou úplně přesné, i když koeficienty jsou přijatelné. Lepších výsledků bychom mohli dosáhnout například změnou algoritmu nebo volbou jiných parametrů modelu.

Text kapitoly 6.2 čerpá z [10], [44].

7 ANALÝZA DATABÁZE FILMŮ A PREDIKCE HODNOCENÍ

Pro praktickou část své práce jsem si vybral jako zdroj dat databázi filmů IMDb [32], která pro nekomerční použití poskytuje zdrojová data v textových souborech zdarma. Tato data bylo nutné upravit do strojově čitelné podoby, protože data stažená ze serverů IMDb [33] nejsou v žádné ze standardních strojově čitelných podob, ale jen jako běžný text. K úpravě dat jsem použil nástroj JMDB [34], který stažené soubory převedl do vlastní lokální databáze MySQL. Celý převod trval téměř tři hodiny. Výsledkem je 50 databázových tabulek, 3 187 806 filmů, 3 245 385 herců (2 100 517 mužů a 1 144 868 žen) a mnoho dalších údajů. Celkově se jedná o 81 279 393 záznamů o celkové velikosti 7,05 GB. Data jsou aktuální ze dne 25. 1. 2015.



Obrázek 32 - Databáze filmů IMDb

7.1 Data dostupná z IMDb

Pro předpověď hodnocení filmů bylo použito 22 datasetů. Níže je seznam všech používaných datasetů se základní informací o jejich obsahu.

- Ratings – Dataset obsahuje ID filmu, hodnocení a počet hlasů.
- Movies – ID filmu, jméno filmu a rok vydání.
- Actors – ID herce, jméno herce a pohlaví.
- Movies2actors – ID filmu, ID herce a informace o tom, jakou postavu herec ve filmu ztvárňuje.
- Directors – ID režiséra a jméno režiséra.
- Movies2directors – ID filmu a ID režiséra.
- Composers – ID skladatele, jméno skladatele.

- Movies2composers – ID filmu a ID skladatele.
- Costdesigners – ID návrháře kostýmů a jeho jméno.
- Movies2costdes – ID filmu a ID návrháře kostýmů.
- Countries – ID filmu a jméno země, ve které se film natáčel.
- Distributors – ID filmu a jméno distributora.
- Editors – ID editora a jeho jméno.
- Movies2editors – ID filmu a ID editora.
- Genres – ID filmu a jeho žánr.
- Language – ID filmu a jeho původní jazyk.
- Locations – ID filmu a jméno konkrétní lokace, ve které se film natáčel.
- Prodcompanies – ID filmu a jméno produkční společnosti.
- Producers – ID producenta a jeho jméno.
- Movies2producers – ID filmu a ID producenta.
- Writers – ID spisovatele a jeho jméno.
- Movies2writers – ID filmu a ID spisovatele.

7.2 Tvorba modelu v Microsoft Azure

Ve strojovém učení je hlavním problémem výběr správného algoritmu a sady parametrů, které nejlépe popisují danou problematiku.

U výběru algoritmu je zřejmé, že se musí jednat o algoritmus založený na regresi. Cílem je predikce hodnocení filmu, tedy hodnota reálné funkce. Microsoft Azure ML nabízí 8 modulů pro regresi:

1. **Bayesian Linear Regression** – Bayesovský přístup používá lineární regresi podpořenou dalšími informacemi ve formě předchozího rozdělení pravděpodobnosti. Předchozí informace o parametrech je kombinovaná s pravděpodobnostní funkcí pro generování odhadů parametrů. [36]
2. **Boosted Decision Tree Regression** – Boosting je jednou z několika klasických metod pro vytváření souboru matic. V této metodě jsou stromy vytvářeny stupňovitě s využitím ztrátové funkce k měření chyby v každém kroku a opravě pro následující kroky. Vytvářena je stupňovitá série stromů a poté je vybrán optimální strom použi-

tím libovolné diferencovatelné ztrátové funkce. V Azure ML Studiu Boosted Decision Trees používají efektivní implementaci MART gradient boosting algoritmu. [37]

3. **Decision Forest Regression** – Rozhodovací stromy jsou neparametrické modely, které provádí sekvence jednoduchých testů pro každou instanci a prochází binární stromovou strukturu dat, dokud není dosaženo listového (rozhodovacího) uzlu. Regresní model se skládá z matice rozhodovacích stromů. [38]
4. **Fast Forest Quantile Regression** – Nejjednodušší definicí kvantilu je hodnota, která rozděluje data na podobně velké skupiny a hodnota kvantilů určuje hranici mezi jednotlivými skupinami. Kvantilová regrese se používá v případě, kdy chceme lépe rozumět rozdělení předpovídaných hodnot. Pokud chceme předpovídat rozsah nebo rozdělení předpovídaných hodnot, použijeme techniky Bayesovské regrese nebo kvantilové regrese. [39]
5. **Linear Regression** – Základním smyslem regrese je přizpůsobení modelu cíli, který je reprezentovaný jako číselný vektor. Lineární regrese se používá v případě, kdy chceme získat velmi jednoduchý model pro základní predikci. Lineární regrese má tendenci správně pracovat na vysoce dimenzionálních datech – řídké datasety postrádají komplexnost. V Azure ML Studiu se lineární regrese používá při řešení více lineárních regresních problémů, ve kterých se nachází jedna závislá proměnná, která má více než jednoho prediktora. [40]
6. **Neural Network Regression** – Neuronové sítě jsou známé hlavně pro deep learning a modelování komplexních problémů jako například rozpoznávání obrazu. Neuronové sítě však mohou být snadno adaptovány i pro regresní problémy. Každá třída statistických modelů může být nazývána neuronovou sítí, pokud používá adaptivní váhy a může aproximovat nelineární funkce svých vstupů. Proto je vhodná pro problémy, kde běžné regresní modely nejsou vhodné. Jedná se o techniku učení s učitelem. [41]
7. **Ordinal Regression** – Ordinální regrese je regresní model, ve kterém cílové hodnoty mají přirozené řazení. Přirozené řazení čísel se používá pro hodnocení. [42]
8. **Poisson Regression** – Poissonova regrese je speciálním typem regresní analýzy, která se typicky používá pro počty modelu. Například modelování nachlazení v souvislosti s létáním letadlem nebo stanovení počtu tísňových volání během nějaké akce apod. Z poskytnutého trénovacího datasetu Poissonova regrese zkouší najít optimální hodnoty maximalizací záznamu pravděpodobnosti jednotlivých parametrů vzhledem

ke vstupům. Pravděpodobnost parametrů je pravděpodobností, s jakou byla trénovací data navzorkována z rozdělení s těmito parametry.

7.3 Předzpracování dat

Původní datasety byly před použitím ještě dále upraveny, aby jednak byla zmenšena jejich velikost, ale také zefektivněno jejich použití. Všechny datasety byly upraveny tím způsobem, že došlo k jejich propojení s hodnocením. Jednotlivé položky datasetů pak byly pomocí SQL dotazu nahrazeny průměrným hodnocením filmů, se kterými jsou spojeny. Konkrétně například u tabulky herců byli herci nahrazeni průměrným hodnocením filmů, ve kterých se daný herec objevil. Tím bylo docíleno jednak odstranění nežádoucích dlouhých textových řetězců popisujících danou položku (například jméno u herce), ale hlavně rovnou k dané položce bylo přiřazeno její průměrné hodnocení ze všech filmů.

view as

actorid	actor_avg
4	6.708333
5	6.5
20	8.5
23	4.2
28	6.338462
31	6.2

Obrázek 33 - Průměrné hodnocení herců podle hodnocení filmů

Dalším krokem byla agregace těchto hodnot do jedné. U herců bylo vypočítáno ještě průměrné hodnocení všech herců, kteří se v daném filmu objevili, které bylo následně propojeno s daným filmem opět pomocí SQL dotazu. Kromě průměru je zaznamenáváno i hodnocení herce s nejlepším hodnocením a herce s nejhorším hodnocením.

view as

movieid	best_actor	avg_actor	worst_actor
3	7.517857	6.809298	6.2
4	7.233333	6.611594	6.2
5	7.119167	6.252374	4.929525
6	6.977778	6.541545	6.242857
7	7.108	6.356439	4.9
8	6.977778	6.513302	6.2
9	7.555405	6.83439	6.2
10	6.977778	6.541594	6.2

Obrázek 34 - Hodnocení herců u jednotlivých filmů

Stejný postup je aplikován i na všechny ostatní datasety. Díky tomu došlo k rozdělení problému předpovědi hodnocení na menší podproblémy podle dílčích datasetů a tedy snížení nároků na výsledný výpočet, navíc tím bylo eliminováno velké množství duplicitních hodnot, protože například u herců není nutné evidovat všechny filmy, ve kterých se objevil, ale stačí evidovat jen jednu důležitou hodnotu průměrného hodnocení všech filmů, ve kterých se daný herec objevil, a s touto hodnotou pak dále pracovat. U tabulky, která eviduje propojení herců s filmy, díky tomu došlo k redukci z původních 23 576 920 záznamů na výsledných 2 169 889.

Základem pro všechny pokusy je model, ve kterém je propojeno hodnocení filmů (dataset ratings) s filmy (dataset movies). Data jsou rozdělena v poměru 70 % trénovací a 30 % testovací.

Při výběru vhodných datasetů se vyskytl problém, že data poskytnutá z IMDb nejsou kompletní. Velmi mnoho filmů nemá vyplněné všechny informace, tudíž při propojení všech datasetů vznikla řídká matice. Při výběru vhodných datasetů tedy bylo nutné brát ohled nejen na korelaci dat s hodnocením filmu, ale také na počet hodnot, aby kvůli malému počtu hodnot nedošlo k odstranění mnoha řádků a tedy zmenšení trénovacího a testovacího datasetu. Protože i když filmů celkově je 3 187 806, hodnocených filmů je jen 594 406. Uvedené číslo je tedy jako výchozí pro určování počtu prázdných hodnot u dalších parametrů.

Pro výpočet korelace byla použita funkce Pearsonovy korelace.

Tabulka 3 - Pearsonova korelace a dostupné hodnoty pro jednotlivé parametry

	Herec průměrný	Režisér	Spisovatel	Produkční společnost	Producent	Editor	Skladatel	Herec nejhorší	Distributor	Herec nejlepší	Lokace	Žánr	Země	Jazyk
Korelace	0,7110	0,7022	0,6857	0,6829	0,6421	0,6406	0,6273	0,5987	0,5752	0,5123	0,4698	0,2859	0,2049	0,2005
Dostupné hodnoty	515 574	427 252	288 037	361 382	385 470	331 456	291 379	515 574	310 129	515 574	198 964	296 746	399 674	424 117
Chybějící hodnoty	78 832	167 154	206 369	233 024	208 936	262 950	303 027	78 832	284 277	78 832	395 442	297 660	194 732	170 289
Chybějící hodnoty	13,26 %	28,12 %	34,72 %	39,20 %	35,15 %	44,24 %	50,98 %	13,26 %	47,83 %	13,26 %	66,53 %	50,08 %	32,76 %	28,65 %

7.4 Porovnání modelů předpovědi

Vzhledem k tomu, že neexistuje exaktní postup pro výběr parametrů, algoritmu a jeho nastavení, bylo potřeba jednotlivé algoritmy a parametry vyzkoušet a porovnat. Jako odchylka je zde brána hodnota MAE (Mean Absolute Error) a i dále v textu bude pod pojmem odchylka myšlena vždy MAE, nebude-li uvedeno jinak. Dále jsou uvedeny počty hodnot, které měly odchylku (MAE) menší než 0,1, 0,2 a 0,3.

7.4.1 Studie 1: Boosted Decision Tree, parametr průměrných herců

Zvolen algoritmus Boosted Decision Tree ve výchozím nastavení Maximum number of leaves per tree = 20; Minimum number of samples per leaf node = 10; Learning rate = 0,2; Total number of trees constructed = 100 a parametr průměrných herců.

Tabulka 4 - Boosted Decision Tree, parametr průměrných herců

Doba výpočtu	Odchylka	Chyba < 0,1	Chyba < 0,2	Chyba < 0,3
0:02:07	0,6689	11,76 %	23,35 %	34,13 %

7.4.2 Studie 2: Boosted Decision Tree, všechny parametry

Zvolen algoritmus Boosted Decision Tree ve výchozím nastavení Maximum number of leaves per tree = 20; Minimum number of samples per leaf node = 10; Learning rate = 0,2; Total number of trees constructed = 100 a všechny dostupné parametry.

Tabulka 5 - Boosted Decision Tree, všechny parametry

Doba výpočtu	Odchylka	Chyba < 0,1	Chyba < 0,2	Chyba < 0,3
0:00:50	0,5041	14,82 %	28,97 %	41,67 %

7.4.3 Studie 3: Neural Network Regression, parametr průměrných herců

Zvolen algoritmus Neural Network Regression ve výchozím nastavení Trainer Mode = Single Parameter; Hidden layer specification = Fully-connected case; Number of hidden nodes = 100; Learning rate = 0,005; Number of learning iterations = 100; The initial learning weights diameter = 0,1; The momentum = 0; The type of normalizer = Min-Max normalizer a parametr průměrných herců.

Tabulka 6 - Neural Network Regression, parametr průměrných herců

Doba výpočtu	Odchylka	Chyba < 0,1	Chyba < 0,2	Chyba < 0,3
0:02:15	0,6839	11,80 %	23,06 %	33,34 %

7.4.4 Studie 4: Neural Network Regression, všechny parametry

Zvolen algoritmus Neural Network Regression ve výchozím nastavení Trainer Mode = Single Parameter; Hidden layer specification = Fully-connected case; Number of hidden nodes = 100; Learning rate = 0,005; Number of learning iterations = 100; The initial learning weights diameter = 0,1; The momentum = 0; The type of normalizer = Min-Max normalizer a všechny dostupné parametry.

Tabulka 7 - Neural Network Regression, všechny parametry

Doba výpočtu	Odchylka	Chyba < 0,1	Chyba < 0,2	Chyba < 0,3
0:01:20	0,5574	13,12 %	25,71 %	36,99 %

7.4.5 Studie 5: Decision Forest Regression, parametr průměrných herců

Zvolen algoritmus Decision Forest Regression ve výchozím nastavení Resampling method = Bagging; Number of decision trees = 8; Maximum depth of the decision trees = 32; Number of random splits per node = 128; Minimum number of samples per leaf node = 1 a parametr průměrných herců.

Tabulka 8 - Decision Forest Regression, parametr průměrných herců

Doba výpočtu	Odchylka	Chyba < 0,1	Chyba < 0,2	Chyba < 0,3
0:03:25	0,7072	11,80 %	22,68 %	32,58 %

7.4.6 Studie 6: Decision Forest Regression, všechny parametry

Zvolen algoritmus Decision Forest Regression ve výchozím nastavení Resampling method = Bagging; Number of decision trees = 8; Maximum depth of the decision trees = 32; Number of random splits per node = 128; Minimum number of samples per leaf node = 1 a všechny dostupné parametry.

Tabulka 9 - Decision Forest Regression, všechny parametry

Doba výpočtu	Odchylka	Chyba < 0,1	Chyba < 0,2	Chyba < 0,3
0:01:15	0,5336	15,11 %	28,86 %	40,99 %

7.4.7 Studie 7: Linear Regression, parametr průměrných herců

Zvolen algoritmus Linear Regression ve výchozím nastavení Solution method = Ordinary Least Squares; L2 regularization weight = 0,001 a parametr průměrných herců.

Tabulka 10 - Linear Regression, parametr průměrných herců

Doba výpočtu	Odchylka	Chyba < 0,1	Chyba < 0,2	Chyba < 0,3
0:01:08	0,6954	11,42 %	22,43 %	32,71 %

7.4.8 Studie 8: Linear Regression, všechny parametry

Zvolen algoritmus Linear Regression ve výchozím nastavení Solution method = Ordinary Least Squares; L2 regularization weight = 0,001 a všechny dostupné parametry.

Tabulka 11 - Linear Regression, všechny parametry

Doba výpočtu	Odchylka	Chyba < 0,1	Chyba < 0,2	Chyba < 0,3
0:01:14	0,5859	12,20 %	23,67 %	34,82 %

Vzhledem k výše uvedeným porovnáním jsem zvolil jako nejvhodnější algoritmus Boosted Decision Tree, který se jeví jako nejrychlejší a přitom nejpřesnější. Předpověď jsem dále vylepšoval úpravou parametrů a nastavení algoritmu.

7.4.9 Studie 9: Boosted Decision Tree, podstatná korelace

Zvolen algoritmus Boosted Decision Tree ve výchozím nastavení Maximum number of leaves per tree = 20; Minimum number of samples per leaf node = 10; Learning rate = 0,2; Total number of trees constructed = 100 a parametry zvoleny dle korelace tak, že byly vybrány pouze parametry s podstatnou korelací s hodnocením filmu (korelace větší než 0,5).

Tabulka 12 - Boosted Decision Tree, podstatná korelace

Doba výpočtu	Odchylka	Chyba < 0,1	Chyba < 0,2	Chyba < 0,3
0:01:12	0,4670	16,90 %	32,28 %	45,22 %

7.4.10 Studie 10: Boosted Decision Tree, podstatná korelace, experimentální nastavení

Zvolen algoritmus Boosted Decision Tree v experimentálně nalezeném optimálním nastavení Maximum number of leaves per tree = 150; Minimum number of samples per leaf node = 12; Learning rate = 0,04; Total number of trees constructed = 700 a parametry zvoleny dle korelace tak, že byly vybrány pouze parametry s podstatnou korelací s hodnocením filmu (korelace větší než 0,5).

Tabulka 13 - Boosted Decision Tree, podstatná kolerace, experimentální nastavení

Doba výpočtu	Odchylka	Chyba < 0,1	Chyba < 0,2	Chyba < 0,3
0:00:52	0,4365	18,50 %	35,38 %	48,81 %

7.4.11 Studie 11: Boosted Decision Tree, podstatná korelace, dostatek hodnot

Zvolen algoritmus Boosted Decision Tree v experimentálně nalezeném optimálním nastavení Maximum number of leaves per tree = 150; Minimum number of samples per leaf node = 12; Learning rate = 0,04; Total number of trees constructed = 700 a parametry zvoleny dle korelace tak, že byly vybrány pouze parametry s podstatnou korelací s hodnocením filmu (korelace větší než 0,5). Byly odstraněny ty parametry, které měly více než 40 % chybějících hodnot.

Tabulka 14 - Boosted Decision Tree, podstatná korelace, dostatek hodnot

Doba výpočtu	Odchylka	Chyba < 0,1	Chyba < 0,2	Chyba < 0,3
0:02:36	0,4676	18,70 %	34,21 %	47,05 %

7.4.12 Studie 12: Boosted Decision Tree, experimentálně zjištěné parametry i nastavení

Zvolen algoritmus Boosted Decision Tree v experimentálně nalezeném optimálním nastavení Maximum number of leaves per tree = 150; Minimum number of samples per leaf node = 12; Learning rate = 0,04; Total number of trees constructed = 700 a experimentálně zvolenými optimálními parametry – všechny parametry kromě žánru, lokace a země.

Tabulka 15 - Boosted Decision Tree, experimentálně zjištěné parametry i nastavení

Doba výpočtu	Odchylka	Chyba < 0,1	Chyba < 0,2	Chyba < 0,3
0:00:50	0,4319	19,29 %	35,80 %	49,35 %

7.4.13 Studie 13: Boosted Decision Tree, experimentálně zjištěné parametry, nastavení dle Sweep Parameters

Zvolen algoritmus Boosted Decision Tree v optimálním nastavení dle modulu Sweep Parameters [45] s hodnotou Random sweep = 200. Optimální nalezené nastavení je Maximum number of leaves per tree = 540; Minimum number of samples per leaf node = 25; Learning rate = 0,0108; Total number of trees constructed = 2107 a experimentálně zvolenými optimálními parametry – všechny parametry kromě žánru, lokace a země. Nalezení těchto nastavení trvalo 9 hodin, 12 minut a 18 vteřin. Zvýšení hodnoty Random sweep na 400 přineslo stejný výsledek optimálního nastavení a výpočet trval 19 hodin, 21 minut a 12 vteřin.

Tabulka 16 - Boosted Decision Tree, experimentálně zjištěné parametry, nastavení dle Sweep Parameters

Doba výpočtu	Odchylka	Chyba < 0,1	Chyba < 0,2	Chyba < 0,3
0:03:15	0,4278	19,33 %	36,18 %	49,41 %

7.4.14 Studie 14: Boosted Decision Tree, experimentálně zjištěné parametry, nastavení dle Sweep Parameters, více než 30 hodnocení

Zvolen algoritmus Boosted Decision Tree v optimálním nastavení dle modulu Sweep Parameters s hodnotou Random sweep = 200. Optimální nalezené nastavení je Maximum number of leaves per tree = 540; Minimum number of samples per leaf node = 25; Learning rate = 0,0108; Total number of trees constructed = 2107 a experimentálně zvolenými optimálními parametry – všechny parametry kromě žánru, lokace a země. Některé filmy mají malý počet hodnocení, což se může projevit na zkreslení předpovědi. Ponechal jsem tedy pouze ty filmy, které měly více než 30 hodnocení, jelikož 30 je již možné považovat za statistický vzorek, což se i experimentálním porovnáním potvrdilo. Tím došlo k odstranění 356 966 filmů (medián počtu hodnocení je pouze 20), ale zpřesnění předpovědi.

Tabulka 17 - Boosted Decision Tree, experimentálně zjištěné parametry, nastavení dle Sweep Parameters, více než 30 hodnocení

Doba výpočtu	Odchylka	Chyba < 0,1	Chyba < 0,2	Chyba < 0,3
0:03:16	0,4179	20,15 %	37,27 %	51,43 %

Z výše uvedených studií nejlépe vychází poslední studie číslo 14, ve které byl použit algoritmus Boosted Decision Tree. Parametry modelu byly zjištěny experimentálním způsobem. Tímto způsobem bylo dosaženo lepších výsledků než při použití korelace i ve spojení s odstraněním parametrů, které obsahovaly nejvíce nevyplněných informací. Vzhledem k velkému rozsahu možných hodnot nastavení algoritmu Boosted Decision Tree nebylo reálné vyzkoušet ručně všechny možnosti. Proto byl použit modul Sweep Parameters, který umožňuje postupným zkoušením různého nastavení najít nejoptimálnější konfiguraci. Aby byly eliminovány filmy s malým počtem hodnocení, byly zahrnuty pouze ty filmy, které měly více než 30 hodnocení.

Tabulka 18 - Porovnání algoritmů pro předpověď hodnocení

	Doba výpočtu	Odchylka	Chyba < 0,1	Chyba < 0,2	Chyba < 0,3
Studie 1	0:02:07	0,6689	11,76 %	23,35 %	34,13 %
Studie 2	0:00:50	0,5041	14,82 %	28,97 %	41,67 %
Studie 3	0:02:15	0,6839	11,80 %	23,06 %	33,34 %
Studie 4	0:01:20	0,5574	13,12 %	25,71 %	36,99 %
Studie 5	0:03:25	0,7072	11,80 %	22,68 %	32,58 %
Studie 6	0:01:15	0,5336	15,11 %	28,86 %	40,99 %
Studie 7	0:01:08	0,6954	11,42 %	22,43 %	32,71 %
Studie 8	0:01:14	0,5859	12,20 %	23,67 %	34,82 %
Studie 9	0:01:12	0,4670	16,90 %	32,28 %	45,22 %
Studie 10	0:00:52	0,4365	18,50 %	35,38 %	48,81 %
Studie 11	0:02:36	0,4676	18,70 %	34,21 %	47,05 %
Studie 12	0:00:50	0,4319	19,29 %	35,80 %	49,35 %
Studie 13	0:03:15	0,4278	19,33 %	36,18 %	49,41 %
Studie 14	0:03:16	0,4179	20,15 %	37,27 %	51,43 %

7.5 Zhodnocení výsledků

Při tvorbě modelu bylo nutné vybrat vhodný algoritmus a parametry. Porovnáním dostupných algoritmů pro regresi byl jako nejvhodnější vybrán algoritmus Boosted Decision Tree, který poskytoval nejmenší odchylku a současně nejkratší dobu výpočtu v základním nastavení.

Pro volbu vhodných parametrů bylo použito několik metod. První metodou výběru parametrů byla korelace s hodnocením, kde byly vybrány parametry s korelací vyšší než 0,5. Druhou metodou byl výběr parametrů s alespoň 60 % dostupných hodnot. Poslední metoda byla experimentální výběr parametrů, která se ukázala jako nejvhodnější, protože poskytovala výsledky s nejmenší odchylkou.

Pro nalezení nejvhodnějšího nastavení algoritmu Boosted Decision Tree byl použit modul Sweep Parameters, který experimentálně porovnává výsledky jednotlivých nastavení. Pro tento postup je vhodné ručně najít alespoň přibližné rozsahy, v jakých se jednotlivé položky nastavení algoritmu mohou pohybovat, a tyto rozsahy pro prohledávání zadat pro porovnání a nalezení nejvhodnějších. Pro Maximum number of leaves per tree byl nastaven rozsah 100 až 1500, pro Minimum number of samples per leaf node rozsah 10 až 50, Learning rate 0,1 až 0,001 a Total number of trees constructed 400 až 4000. Aby porovnání netrvalo extrémně dlouho, bylo nastaveno 200 porovnání, což trvalo 9 hodin, 12 minut a 18 vteřin. Při zvýšení počtu porovnání již nedošlo k nalezení lepšího nastavení a doba výpočtu rostla přibližně lineárně.

Protože některé filmy měly malý počet hodnocení, který by mohl zkreslit celkové hodnocení a tedy i učení sítě, byly odstraněny filmy s méně než 30 hodnoceními, čímž došlo i ke zmenšení odchylky.

Jako nejlepší byla vyhodnocena Studie 14, ve které bylo dosaženo nejmenší odchylky 0,4179 od skutečných hodnot. Více než polovina testovacích hodnot měla chybu menší než 0,3, což lze vzhledem k velkému množství chybějících údajů v datasetech považovat za dobrý výsledek. Hodnocení filmů navíc není možné zcela přesně předpovědět, protože i dobré výchozí předpoklady (herecké obsazení, režisér, scénárista, …) mohou přinést film, který se nebude divákům líbit a budou ho tedy špatně hodnotit, a naopak špatné výchozí předpoklady mohou přinést film, které od diváků dostane dobré hodnocení.

ZÁVĚR

Cílem práce bylo vytvoření praktického návodu pro práci s Azure Machine Learning, který je v kapitole 6.2. Dále byly v kapitole 7 demonstrovány možnosti Azure Machine Learning na předpovědi hodnocení filmů.

Jako nejlepší algoritmus pro předpověď hodnocení filmů byl vybrán Boosted Decision Tree, který přinesl nejmenší odchylku předpovědi a současně nejkratší dobu výpočtu. Při porovnání parametrů modelu vyšla nejlépe experimentální metoda výběru, která poskytla lepší výsledky než při výběru parametrů z korelace nebo podle počtu chybějících hodnot. Nastavení algoritmu Boosted Decision Tree bylo nalezeno experimentálním způsobem pomocí dostupného modulu Sweep Parameters. Dále byly odstraněny filmy, které měly méně než 30 hodnocení, aby nedošlo ke zkreslení předpovědi. Takto naučený a otestovaný model přinesl odchylku (Mean Absolute Error) 0,4179 a 51,43 % testovaných hodnot mělo chybu menší než 0,3.

Azure Machine Learning umožňuje publikování modelů na webových stránkách pro využití veřejnosti. Takto publikovaný model pro předpověď hodnocení filmů by mohli využít například uživatelé, kteří plánují navštívit nový film v kině, který ale ještě nemá žádné hodnocení, aby se přesvědčili, jestli lze předpokládat, že daný film bude dobrý z předpovězeného hodnocení. Další využití by mohlo být například pro produkční společnosti nebo studia při rozhodování o natočení například další série seriálu nebo pro herecké obsazení v připravovaném filmu. Aby bylo model možné využít, publikoval jsem jeho upravenou a zjednodušenou (ale plně funkční) verzi do galerie Microsoft Azure [46].

Hlavním cílem práce bylo prozkoumání a zdokumentování možností Microsoft Azure a jeho nástrojů pro strojové učení. Tomuto cíli se věnuje celá praktická část práce.

SEZNAM POUŽITÉ LITERATURY

[1] MICROSOFT CORPORATION. Microsoft's Cloud Platform: Azure [online]. Seattle, 2014 [cit. 2014-10-07]. Dostupné z: http://azure.microsoft.com/en-us/

[2] MICROSOFT CORPORATION. TechNet Blog CZ/SK: Microsoft Azure [online]. Praha, 2014 [cit. 2014-10-07]. Dostupné z: http://blogs.technet.com/b/technet-czsk/p/azure.aspx

[3] MICROSOFT CORPORATION. TechNet Blogs: Machine Learning [online]. Seattle, 2014 [cit. 2014-10-07]. Dostupné z: http://blogs.technet.com/b/machinelearning/

[4] Creating an intelligent "sandbox" for coordinated malware eradication. MICROSOFT CORPORATION. Malware Protection Center [online]. Seattle, 2014 [cit. 2014-10-07]. Dostupné z: http://blogs.technet.com/b/mmpc/archive/2014/03/31/creating-an-intelligent-sandbox-for-coordinated-malware-eradication.aspx

[5] Recommendations Everywhere. MICROSOFT CORPORATION. Machine Learning Blog [online]. Seattle, 2014 [cit. 2014-10-09]. Dostupné z: http://blogs.technet.com/b/machinelearning/archive/2014/07/09/recommendations-everywhere.aspx

[6] HECKERMAN, David. A Tutorial on Learning With Bayesian Networks. MICROSOFT CORPORATION. Microsoft Research [online]. Seattle, 1995 [cit. 2014-12-04]. Dostupné z: http://research.microsoft.com/apps/pubs/?id=69588

[7] ŠENOVSKÝ, Pavel. Modelování rozhodovacích procesů [online]. 2. vydání. Ostrava: Vysoká škola báňská – Technická univerzita Ostrava, 2009, s. 5-6 [cit. 2014-12-04].

[8] JJ Food Service: Food Delivery Service Uses Machine Learning to Revolutionize Customer Service. MICROSOFT CORPORATION. Customer Stories [online]. Seattle, 2014 [cit. 2014-12-04]. Dostupné z: https://customers.microsoft.com/Pages/CustomerStory.aspx?recid=12792

[9] Carnegie Mellon University: Carnegie Mellon Sees a Way to Cut Energy Use by 20 Percent with Cloud Machine Learning Solution. MICROSOFT CORPORATION. Customer Stories [online]. Seattle, 2014 [cit. 2014-12-04]. Dostupné z: https://customers.microsoft.com/Pages/CustomerStory.aspx?recid=8576

[10] Create a simple experiment in Azure Machine Learning Studio. MICROSOFT CORPORATION. Microsoft Azure [online]. Seattle, 2015 [cit. 2015-01-16]. Dostupné z: http://azure.microsoft.com/en-us/documentation/articles/machine-learning-create-experiment/

[11] ALFELD, Peter. Random Number Generators. The University of Utah [online]. 2011 [cit. 2015-02-04]. Dostupné z: http://www.math.utah.edu/~pa/Random/Random.html

[12] Meet Cortana for Windows Phone. MICROSOFT CORPORATION. Windows Phone [online]. Seattle, 2015 [cit. 2015-01-21]. Dostupné z: http://www.windowsphone.com/en-us/how-to/wp8/cortana/meet-cortana

[13] Siri. APPLE INC. Apple: iOS [online]. Cupertino, 2015 [cit. 2015-01-21]. Dostupné z: https://www.apple.com/ios/siri/

[14] What is cloud computing?. SALESFORCE.COM, Inc. CRM & Cloud Computing To Grow Your Business [online]. San Francisco, CA, United States, 2000-2015 [cit. 2015-02-04]. Dostupné z: http://www.salesforce.com/cloudcomputing/

[15] What Is Big Data?: Big Data. What it is & why it matters. SAS INSTITUTE INC. Business Analytics and Business Intelligence Software [online]. 2015 [cit. 2015-02-04]. Dostupné z: http://www.sas.com/en_us/insights/big-data/what-is-big-data.html

[16] HINTON, Geoffrey E. A tutorial on Deep Learning. In: VideoLectures.NET [online]. University of Toronto, 2009-09-15 [cit. 2015-02-04]. Dostupné z: http://videolectures.net/jul09_hinton_deeplearn/

[17] ROE, Byron P., Hai-Jun YANG a Ji ZHU. Boosted Decision Trees, a Powerful Event Classifier. In: Boosted Decision Trees, a Powerful Event Classifier [online]. University of Michigan, 2008 [cit. 2015-01-21]. Dostupné z: http://www.curtis-meyer.com/articles/bkgrnd/phystat05-proc.pdf

[18] R FOUNDATION. The R Foundation for Statistical Computing [online]. 2014 [cit. 2015-01-21]. Dostupné z: http://www.r-project.org

[19] Vowpal Wabbit. GitHub [online]. 2010, 2014-10-21 [cit. 2015-01-21]. Dostupné z: https://github.com/JohnLangford/vowpal_wabbit/wiki

[20] SZEGEDY, Christian, Alexander TOSHEV a Dumitru ERHAN. Deep Neural Networks for Object Detection. In: Deep Neural Networks for Object Detection [online]. 2013 [cit. 2015-01-21]. Dostupné z: http://papers.nips.cc/paper/5207-deep-neural-networks-for-object-detection.pdf

[21] SHOTTON, Jamie, Toby SHARP, Pushmeet KOHLI, Sebastian NOWOZIN, John WINN a Antonio CRIMINISI. Decision Jungles: Compact and Rich Models for Classification. In: Decision Jungles [online]. 2013 [cit. 2015-01-21]. Dostupné z: http://research.microsoft.com/pubs/205439/DecisionJunglesNIPS2013.pdf

[22] RIFKIN, Ryan. Multiclass Classification. In: Massachusetts Institute of Technology: Statistical Learning Theory and Applications [online]. Cambridge, MA, USA, 2008-02-25 [cit. 2015-02-04]. Dostupné z: http://www.mit.edu/~9.520/spring09/Classes/multiclass.pdf

[23] DAUMÉ, Hal. A Course in Machine Learning: BEYOND BINARY CLASSIFICATION. In: [online]. 0.8, 2012-08 [cit. 2015-02-04]. Dostupné z: http://ciml.info/dl/v0_8/ciml-v0_8-ch05.pdf

[24] An Introduction to Regression Analysis. In: SYKES, Alan O. [online]. [cit. 2015-01-21]. Dostupné z: http://www.law.uchicago.edu/files/files/20.Sykes_.Regression.pdf

[25] Cluster Analysis: How To Group Objects Into Similar Categories. STATSOFT INC. StatSoft [online]. [cit. 2015-01-21]. Dostupné z: http://www.statsoft.com/Textbook/Cluster-Analysis

[26] MIZONOV, Valery a Seth MANHEIM. Azure Table Storage and Windows Azure SQL Database: Compared and Contrasted. MICROSOFT CORPORATION. Microsoft Azure [online]. Seattle, 2014-10-01 [cit. 2015-01-21]. Dostupné z: https://msdn.microsoft.com/en-us/library/azure/jj553018.aspx

[27] How to use Blob Storage from .NET: What is Blob Storage. MICROSOFT CORPORATION. Microsoft Azure [online]. Seattle, 2014-10-11 [cit. 2015-01-21]. Dostupné z: http://azure.microsoft.com/en-us/documentation/articles/storage-dotnet-how-to-use-blobs/#what-is

[28] APACHE SOFTWARE FOUNDATION. Apache Hive TM [online]. 2011 [cit. 2015-01-21]. Dostupné z: https://hive.apache.org/

[29] REFSNES DATA. W3Schools: SQL Tutorial [online]. [cit. 2015-01-21]. Dostupné z: http://www.w3schools.com/sql/

[30] MALÝ, Martin. REST: architektura pro webové API. Zdroják: o tvorbě webových stránek a aplikací [online]. 2009 [cit. 2015-01-21]. Dostupné z: http://www.zdrojak.cz/clanky/rest-architektura-pro-webove-api/

[31] APACHE SOFTWARE FOUNDATION. Apache Hadoop [online]. 2014-12-12 [cit. 2015-01-21]. Dostupné z: http://hadoop.apache.org/

[32] IMDB.COM, Inc. IMDb [online]. 1990-2015 [cit. 2015-01-27]. Dostupné z: http://www.imdb.com/

[33] Alternative Interfaces. IMDB.COM, Inc. IMDb [online]. 1990-2015 [cit. 2015-01-27]. Dostupné z: http://www.imdb.com/interfaces

[34] FREESE, Uwe a Juergen ULBTS. Java Movie Database [online]. 2000, 2014-03-12 [cit. 2015-01-27]. Dostupné z: http://www.jmdb.de/

[35] Machine Learning Pricing. MICROSOFT CORPORATION. Microsoft Azure [online]. 2015 [cit. 2015-03-04]. Dostupné z: http://azure.microsoft.com/en-us/pricing/details/machine-learning/

[36] Bayesian Linear Regression. MICROSOFT CORPORATION. Microsoft Azure [online]. Seattle, 2015 [cit. 2015-03-09]. Dostupné z: https://msdn.microsoft.com/en-us/library/azure/dn906022.aspx

[37] Boosted Decision Tree Regression. MICROSOFT CORPORATION. Microsoft Azure [online]. Seattle, 2015 [cit. 2015-03-09]. Dostupné z: https://msdn.microsoft.com/en-us/library/azure/dn905801.aspx

[38] Decision Forest Regression. MICROSOFT CORPORATION. Microsoft Azure [online]. Seattle, 2015 [cit. 2015-03-09]. Dostupné z: https://msdn.microsoft.com/en-us/library/azure/dn905862.aspx

[39] Fast Forest Quantile Regression. MICROSOFT CORPORATION. Microsoft Azure [online]. Seattle, 2015 [cit. 2015-03-09]. Dostupné z: https://msdn.microsoft.com/en-us/library/azure/dn913093.aspx

[40] Linear Regression. MICROSOFT CORPORATION. Microsoft Azure [online]. Seattle, 2015 [cit. 2015-03-09]. Dostupné z: https://msdn.microsoft.com/en-us/library/azure/dn905978.aspx

[41] Neural Network Regression. MICROSOFT CORPORATION. Microsoft Azure [online]. Seattle, 2015 [cit. 2015-03-09]. Dostupné z: https://msdn.microsoft.com/en-us/library/azure/dn905924.aspx

[42] Ordinal Regression. MICROSOFT CORPORATION. Microsoft Azure [online]. Seattle, 2015 [cit. 2015-03-09]. Dostupné z: https://msdn.microsoft.com/en-us/library/azure/dn906029.aspx

[43] Poisson Regression. MICROSOFT CORPORATION. Microsoft Azure [online]. Seattle, 2015 [cit. 2015-03-09]. Dostupné z: https://msdn.microsoft.com/en-us/library/azure/dn905988.aspx

[44] ELSTON, Stephen F. Data Science in the Cloud with Microsoft Azure Machine Learning and R [online]. First Release. 1005 Gravenstein Highway North, Sebastopol, CA: O'Reilly Media, Inc., 2015 [cit. 2015-03-09]. ISBN 978-1-491-91959-0. Dostupné z: https://azureinfo.microsoft.com/CO-Azure-CNTNT-FY15-02Feb-Data-Science-in-the-Cloud.html?ls=Media&lsd=Oreilly

[45] Sweep Parameters. MICROSOFT CORPORATION. Microsoft Azure [online]. Seattle, 2015 [cit. 2015-04-07]. Dostupné z: https://msdn.microsoft.com/library/azure/038d91b6-c2f2-42a1-9215-1f2c20ed1b40

[46] Movie rating prediction. 2015. Microsoft Azure Machine Learning Gallery [online]. Seattle [cit. 2015-05-05]. Dostupné z: http://gallery.azureml.net/Experiment/4dcd764f36314692878bae9242c9f9bb

SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK

API	Application Programming Interface, rozhraní pro přístup aplikací
CDN	Content Delivery Network, velký distribuovaný systém serverů určený pro doručování datově náročného obsahu uživatelům po celém světě.
FDA	U. S. Food and Drug Administration, česky Úřad pro kontrolu potravin a léčiv, který je vládní agenturou Spojených států amerických zodpovídající za kontrolu a regulaci potravin, doplňků stravy, léčiv, kosmetických přípravků, lékařských přístrojů a biofarmaceutických a krevních produktů v USA.
IaaS	Infrastructure as a Service, jedna z možností distribuce služeb v cloud computingu, kdy poskytovatel nabízí výpočetní infrastrukturu v dojednané konfiguraci (servery, datová úložiště, síťové prvky a další) jako službu zákazníkům za pravidelný poplatek.
ML	Machine Learning
PaaS	Platform as a Service, oproti [1] se tento typ distribuce služeb liší v tom, že poskytovatel zajišťuje i operační systém celého řešení včetně potřebných nadstaveb, díky čemuž zákazníkovi odpadá nutnost starat se o nasazení, správu a aktualizaci software.
REST	Representational State Transfer, architektura rozhraní pro distribuované prostředí
SLA	Zkratka SLA označuje smlouvu sjednanou mezi poskytovatelem služby a jejím konzumentem. Tato smlouva vymezuje vlastnosti a parametry poskytované služby.
SQL	Structured Query Language, standardizovaný strukturovaný dotazovací jazyk

SEZNAM OBRÁZKŮ

Obrázek 1 - Matice uživatelů, položek a hodnocení. Převzato z [5] 13
Obrázek 2 - Dvojrozměrný latentní prostor parametrů. Převzato z [5] 14
Obrázek 3 - Microsoft Azure Preview portál.....29
Obrázek 4 - Microsoft Azure portál.....30
Obrázek 5 - Vytvoření nové pracovní plochy Machine Learning 31
Obrázek 6 - Workspace v Microsoft Azure 32
Obrázek 7 - Testovací experiment v ML Studiu.....33
Obrázek 8 - Vizualizace dat v ML Studiu 34
Obrázek 9 - Historie běhu experimentu 34
Obrázek 10 - Tvorba nového experimentu 36
Obrázek 11 - Pojmenování experimentu 36
Obrázek 12 - Vybrání dat pro experiment 36
Obrázek 13 - Volba u dat.....37
Obrázek 14 - Vizualizace dat z datasetu 37
Obrázek 15 - Připojení modulu pro odstranění sloupců 38
Obrázek 16 - Výběr sloupců pro odstranění 39
Obrázek 17 - Výběr sloupců pro odstranění 39
Obrázek 18 - Přidání komentáře k modulu 40
Obrázek 19 - Přidání modulu pro odstranění řádků s prázdnými hodnotami 40
Obrázek 20 - Odstranění řádků s prázdnými hodnotami 41
Obrázek 21 - Dokončený experiment 42
Obrázek 22 - Předpřipravený dataset.....43
Obrázek 23 - Výběr parametrů pro předpověď.....44
Obrázek 24 - Rozdělení hodnot na trénovací a testovací.....45
Obrázek 25 - Přidání modulu lineární regrese 46
Obrázek 26 - Výběr dat, které má model předpovídat.....46
Obrázek 27 - Natrénovaný regresní model pro tvorbu predikcí 46
Obrázek 28 - Ohodnocení modelu 47
Obrázek 29 - Výsledek ohodnocení modelu 48
Obrázek 30 - Vyhodnocení modelu 48
Obrázek 31 - Statistické informace o modelu.....49
Obrázek 32 - Databáze filmů IMDb 50

Obrázek 33 - Průměrné hodnocení herců podle hodnocení filmů 53
Obrázek 34 - Hodnocení herců u jednotlivých filmů 54

SEZNAM TABULEK

Tabulka 1 - Ceny Azure Machine Learning 25
Tabulka 2 - Porovnání parametrů služeb Azure Machine Learning 25
Tabulka 3 - Pearsonova korelace a dostupné hodnoty pro jednotlivé parametry 55
Tabulka 4 - Boosted Decision Tree, parametr průměrných herců 56
Tabulka 5 - Boosted Decision Tree, všechny parametry 56
Tabulka 6 - Neural Network Regression, parametr průměrných herců 57
Tabulka 7 - Neural Network Regression, všechny parametry 57
Tabulka 8 - Decision Forest Regression, parametr průměrných herců.....57
Tabulka 9 - Decision Forest Regression, všechny parametry 58
Tabulka 10 - Linear Regression, parametr průměrných herců 58
Tabulka 11 - Linear Regression, všechny parametry 58
Tabulka 12 - Boosted Decision Tree, podstatná korelace 59
Tabulka 13 - Boosted Decision Tree, podstatná kolerace, experimentální nastavení 59
Tabulka 14 - Boosted Decision Tree, podstatná korelace, dostatek hodnot 59
Tabulka 15 - Boosted Decision Tree, experimentálně zjištěné parametry i nastavení.....60
Tabulka 16 - Boosted Decision Tree, experimentálně zjištěné parametry, nastavení dle Sweep Parameters.....60
Tabulka 17 - Boosted Decision Tree, experimentálně zjištěné parametry, nastavení dle Sweep Parameters, více než 30 hodnocení.....61
Tabulka 18 - Porovnání algoritmů pro předpověď hodnocení.....62