

**SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
FAKULTA ELEKTROTECHNIKY A INFORMATIKY**

Machine Learning and cloud AZURE

Učebný text

Oblasť strojové učenie - matematika a umelá inteligencia

Obsah

Úvod.....	1
1 Machine learning & Artificial intelligence & Natural language processing.....	2
1.1 Pracovný postup pri práci so strojovým učením	2
2 Azure – strojové učenie.....	3
2.1 Inicializácia modelu.....	3
2.2 Trénovanie	10
2.3 Skórovanie	10
2.4 Hodnotenie.....	10
3 AWS – strojové učenie.....	11
4 Predikcia ceny auta – príklad strojového učenia v Azure	12
Záver.....	18
Zdroje	19

Zoznam obrázkov

Obr. 1 Boosted decision tree – algoritmus pre regresiu	4
Obr. 2 K-means - algoritmus pre zhlukovanie	6
Obr. 3 Azure - rozdelenie algoritmov pre strojové učenie	9
Obr. 4 Príklad - model strojového učenia pre predikciu ceny auta	12
Obr. 5 Príklad - údaje o jednotlivých automobiloch	13
Obr. 6 Príklad - zvolené vlastnosti automobilu.....	14
Obr. 7 Príklad grafu lineárnej regresie	14
Obr. 8 Príklad - skórovanie - známa hodnota vs predikovaná hodnota	15
Obr. 9 Príklad - Vyhodnotenie modelu	15
Obr. 10 Príklad – Webový servis s natrénovaným modelom.....	16
Obr. 11 Príklad – Možnosť otestovať webový servis predikcie ceny auta	17

Zoznam skratiek a značiek

AI	Umelá inteligencia
AWS	Amazon web servisy
ML	Strojové učenie
NLP	Spracovanie prirodzeného jazyka

Zoznam tabuliek

Tab. 1 Azure – algoritmy pre regresiu	5
Tab. 2 Azure – algoritmus pre zhlukovanie	6
Tab. 3 Azure – algoritmy pre detekciu anomálií.....	7
Tab. 4 Azure – algoritmy pre klasifikáciu.....	8

Úvod

Strojové učenie je v súčasnosti jednou z najhorúcejších tém v oblasti IT. Technológie ako digitálne dvojča, big data, umelá inteligencia, automatizácia a strojové učenie čoraz viac formujú budúcnosť práce. Strojové učenie je špecifický set techník, ktoré umožňujú strojom získavať informácie z údajov a robiť predpovede - predikcie. S vývojom strojového učenia narastá aj efektívnosť robotov, ktoré sa využívajú pri rôznych výrobných činnostiach. Vďaka ML sa dokážu učiť a zefektívňovať výrobné procesy.

Úvodné kapitoly práce sa venujú opisu strojového učenia, ktoré je poskytované v cloud-e od spoločnosti Microsoft na platforme Azure a od spoločnosti Amazon na platforme Amazon Web Services. Primárne sa práca zameriava na strojové učenie nachádzajúce sa v Azure, nakoľko je rozsiahlejšie a ponúka viac algoritmov strojového učenia. Práca sa zameriava na opis jednotlivých krokov, ktoré treba absolvovať pri tvorbe modelu strojového učenia. Záverečná kapitola opisuje proces tvorby modelu strojového učenia pre predikciu ceny automobilu.

1 Machine learning & Artificial intelligence & Natural language processing

Machine Learning (ML) - Strojové učenie je metóda analýzy údajov, ktorá automatizuje budovanie analytického modelu. Ide o odvetvie umelej inteligencie založenej na myšlienke, že systémy sa môžu učiť z dát, identifikovať vzory a na základe vzorov generovať predpovede alebo odporúčania.

Umelá inteligencia (AI) umožňuje strojom učiť sa zo skúseností, prispôbovať sa novým vstupom a vykonávať ľudské úlohy. Väčšina príkladov o AI, o ktorých dnes počujete - od počítačových programov hrajúcich šach až po autonómne vozidlá - sa veľmi spoliehajú na vzdelávanie a počítačové spracovanie prirodzeného jazyka (NLP). Pomocou týchto technológií môžu byť počítače vyškolené na vykonávanie konkrétnych úloh, pričom budú spracovávať veľké množstvo údajov a rozpoznávať v nich vzory.

V stručnosti, spracovanie prirodzeného jazyka (NLP) je určené na pochopenie požiadaviek od používateľa, tak aby systém AI mohol určiť, ako najlepšie vykonať zadanú úlohu [1]. NLP a ML sú podmnožinou AI.

1.1 Pracovný postup pri práci so strojovým učením

Pracovný postup pri práci so strojových učením zahŕňa niekoľko fáz:

- I. Identifikácia metrík a problému, ktorý chceme riešiť.
- II. Nájdenie a čistenie údajov, t. j. príprava vhodných údajov.
- III. Identifikácia najvhodnejších funkcií pre riešenie daného problému, prípadne tvorba nových funkcií.
- IV. Budovanie, vyhodnocovanie a ladenie modelu - modelov.
- V. Používanie modelov na generovanie predpovedí, odporúčaní a iných užitočných poznatkov.

2 Azure – strojové učenie

Azure je cloud computing-ová služba vytvorená spoločnosťou Microsoft slúžiaca na tvorbu, testovanie, nasadzovanie a správu aplikácií a služieb prostredníctvom globálnej siete dátových centier distribuovaných po celom svete. Azure poskytuje softvér ako službu (SaaS), platformu ako službu (PaaS) a infraštruktúru ako službu (IaaS), pričom podporuje mnoho rôznych programovacích jazykov, nástrojov a softvérov. Azure bol oznámený v októbri 2008 s kódovým menom „Project Red Dog“ a oficiálne bol spustený vo februári 2010. Súčasťou Azure je aj Machine Learning Studio, ktoré ponúka sadu nástroj pre prácu so strojovým učením [2].

Práca so strojovým učením v Azure sa skladá zo 4 krokov:

- I. **Inicializácia modelu** – Výber vhodného modelu z veľkej škály prispôsobiteľných algoritmov strojového učenia, vrátane modelov zhľukovania, regresie, klasifikácie a detekcie anomálií.
- II. **Trénovanie** – Poskytnutie údajov nakonfigurovanému modelu, aby sa mohol učiť a vytvoriť štatistiky, ktoré sa dajú využiť pre predikcie.
- III. **Skórovanie** – Vytvorenie predikcie pomocou vyškoleného modelu.
- IV. **Hodnotenie** – Zmeranie presnosti modelu, prípadne porovnanie s výsledkami získanými z iných modelov.

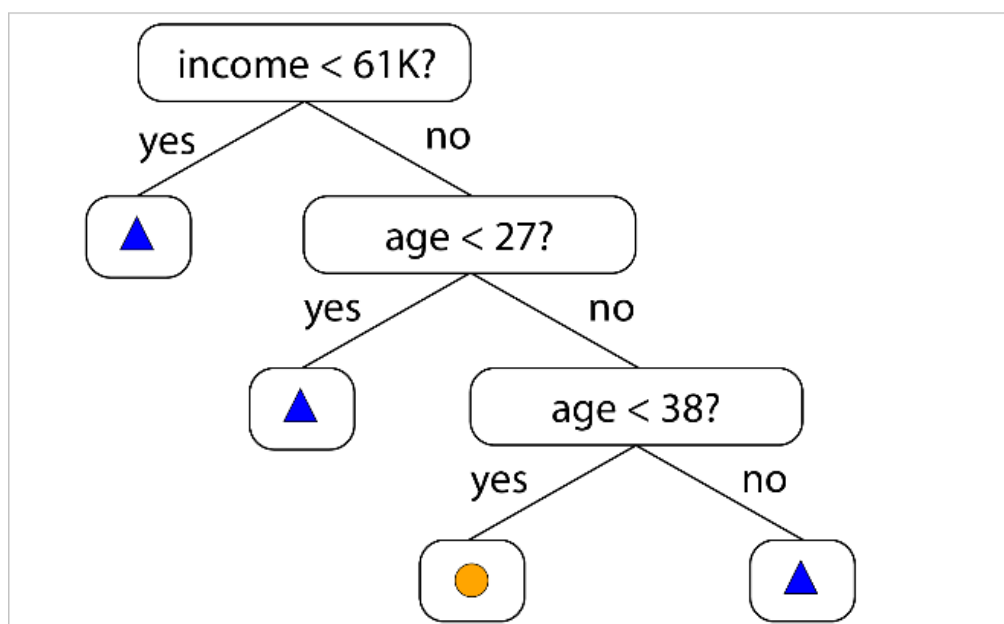
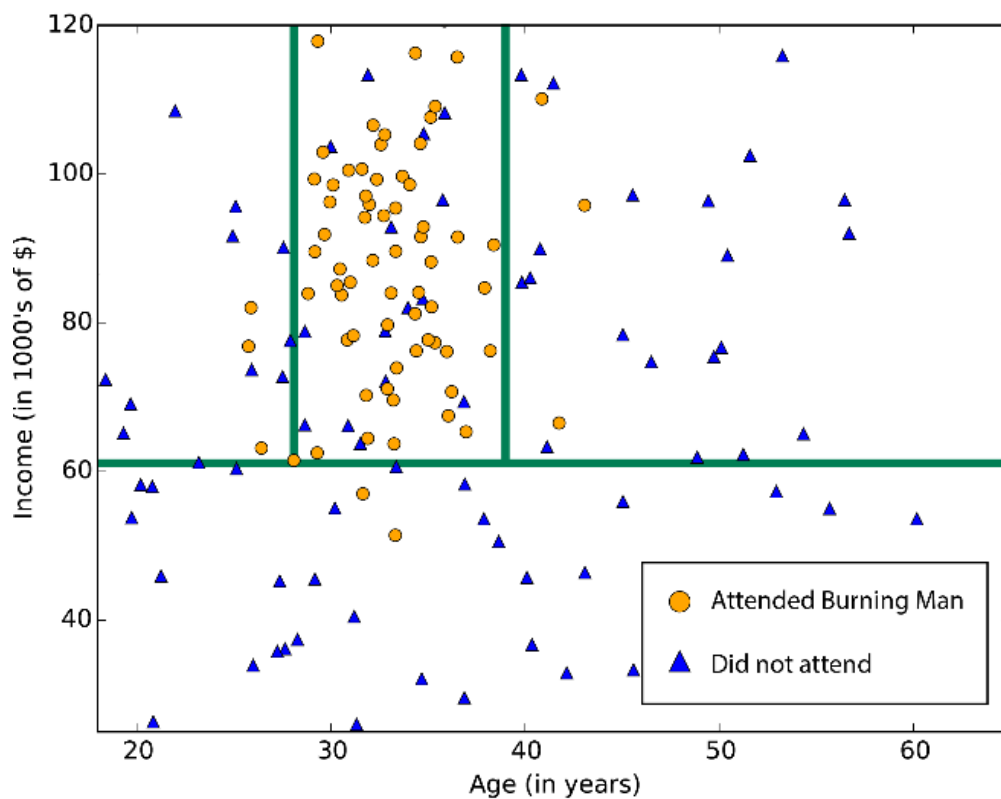
2.1 Inicializácia modelu

Prvou úlohou je identifikovať všeobecnú kategóriu úlohy, ktorú má strojové učenie vykonať a potom vybrať vhodný algoritmus. Poznáme 4 hlavné kategórie modelov, pričom každý model obsahuje moduly, ktoré využívajú rôzne algoritmy (Obr. 1).

Hlavné kategórie modelov:

- **Regresia** – je široko používaná metodológia v oblastiach od inžinierstva až po vzdelávanie. Regresné úlohy sú podporované v mnohých nástrojoch: napríklad Excel poskytuje „What if“ analýzu alebo predpovedanie v priebehu času. Vo všeobecnosti sa regresný algoritmus pokúša naučiť hodnotu funkcie pre konkrétnu inštanciu dát. Regresia môže byť využitá napríklad na [3]:
 - Predpovedanie predajnej ceny domu na základe regionálnych údajov alebo na predpovedanie predajnej ceny domu v budúcnosti.
 - Predpovedanie výšky osoby pomocou funkcie výšky.

- Predpovedanie pravdepodobnosti hospitalizácie človeka na základe hodnôt lekárskeho testov.
- Predpovedanie výkonnosti študentov.
- Predpovedanie odpovedí prieskumu, ktoré respondenti zaškrknú v hranici od 1 -5.
- Predpovedanie poradia pretekárov.



Obr. 1 Boosted decision tree – algoritmus pre regresiu

Nasledujúca tabuľka zobrazuje algoritmy dostupné pre regresiu.

Algorithm	Accuracy	Training time	Linearity	Parameters	Notes
Regression					
linear		●	●	4	
Bayesian linear		○	●	2	
decision forest	●	○		6	
boosted decision tree	●	○		5	Large memory footprint
fast forest quantile	●	○		9	Distributions rather than point predictions
neural network	●			9	Additional customization is possible
Poisson			●	5	Technically log-linear. For predicting counts
ordinal				0	For predicting rank-ordering

Algorithm properties:

- - shows excellent accuracy, fast training times, and the use of linearity
- - shows good accuracy and moderate training times

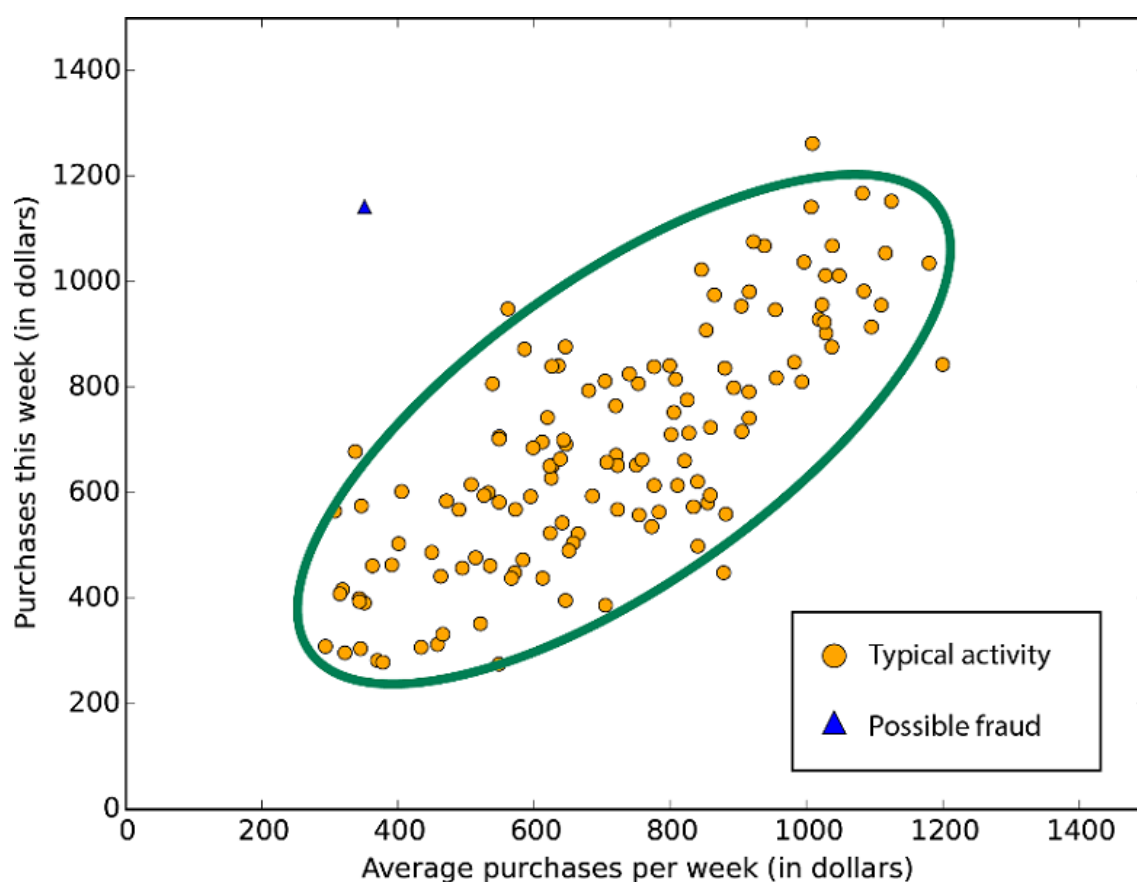
Tab. 1 Azure – algoritmy pre regresiu

- **Zhlukovanie** - v strojovom učení je metóda zoskupovania dátových bodov do podobných klastrov. Tiež sa nazýva aj segmentácia. V priebehu rokov sa vyvinulo množstvo algoritmov zhľukovania, pričom takmer všetky využívajú vlastnosti/rysy jednotlivých položiek na hľadanie podobných položiek. Zhľukovanie sa nazýva technológiou výučby, ktorá nie je pod dohľadom, pretože môže byť použitá na neoznačených údajoch. Neoznačené údaje sú vzorky údajov, ktoré môžeme jednoducho získať zo sveta a neobsahujú žiadne dodatočné údaje (napr. značku, štítok, alebo triedu do ktorej patria) na základe ktorých ich je možné zoskupiť. Zhľukovanie je užitočným prvým krokom na objavovanie nových vzorov a vyžaduje len málo predchádzajúcich poznatkov o tom, ako môžu byť údaje štruktúrované alebo ako spolu údaje súvisia. Často sa používa na preskúmanie údajov pred samotnou analýzou s inými prediktívnymi algoritmi. [4]. Zhľukovanie sa môže využiť napríklad na:
 - Vyhládanie podobných ľudí na základe demografických údajov.
 - Analýzu textu za účelom zoskupenia viet s podobnou témou.

Nasledujúca tabuľka zobrazuje algoritmus dostupný pre zhlukovanie.

Algorithm	Accuracy	Training time	Linearity	Parameters	Notes
Anomaly detection					
K-means		○	●	4	A clustering algorithm
Algorithm properties:					
● - shows excellent accuracy, fast training times, and the use of linearity					
○ - shows good accuracy and moderate training times					

Tab. 2 Azure – algoritmus pre zhlukovanie



Obr. 2 K-means - algoritmus pre zhlukovanie

- **Detekcia anomálií** - zahŕňa mnoho dôležitých úloh v strojovom učení:
 - Identifikácia transakcií, ktoré sú potenciálne podvodné.
 - Detekcia narušenia siete.
 - Nájdenie abnormálnych zoskupení buniek u pacientov.
 - Kontrola hodnôt zadaných do systému.

Pretože anomálie sú zriedkavými udalosťami, môže byť ťažké zhromaždiť reprezentatívnu vzorku dát, ktorá by sa dala použiť. Algoritmy v tabuľke 3 boli špeciálne navrhnuté tak, aby riešili výzvy budovania a trénovania modelov pomocou nevyvážených dátových setov [5].

Algorithm	Accuracy	Training time	Linearity	Parameters	Notes
Anomaly detection					
support vector machine	○	○		2	Especially good for large feature sets
PCA-based anomaly detection		○	●	3	
Algorithm properties:					
● - shows excellent accuracy, fast training times, and the use of linearity					
○ - shows good accuracy and moderate training times					

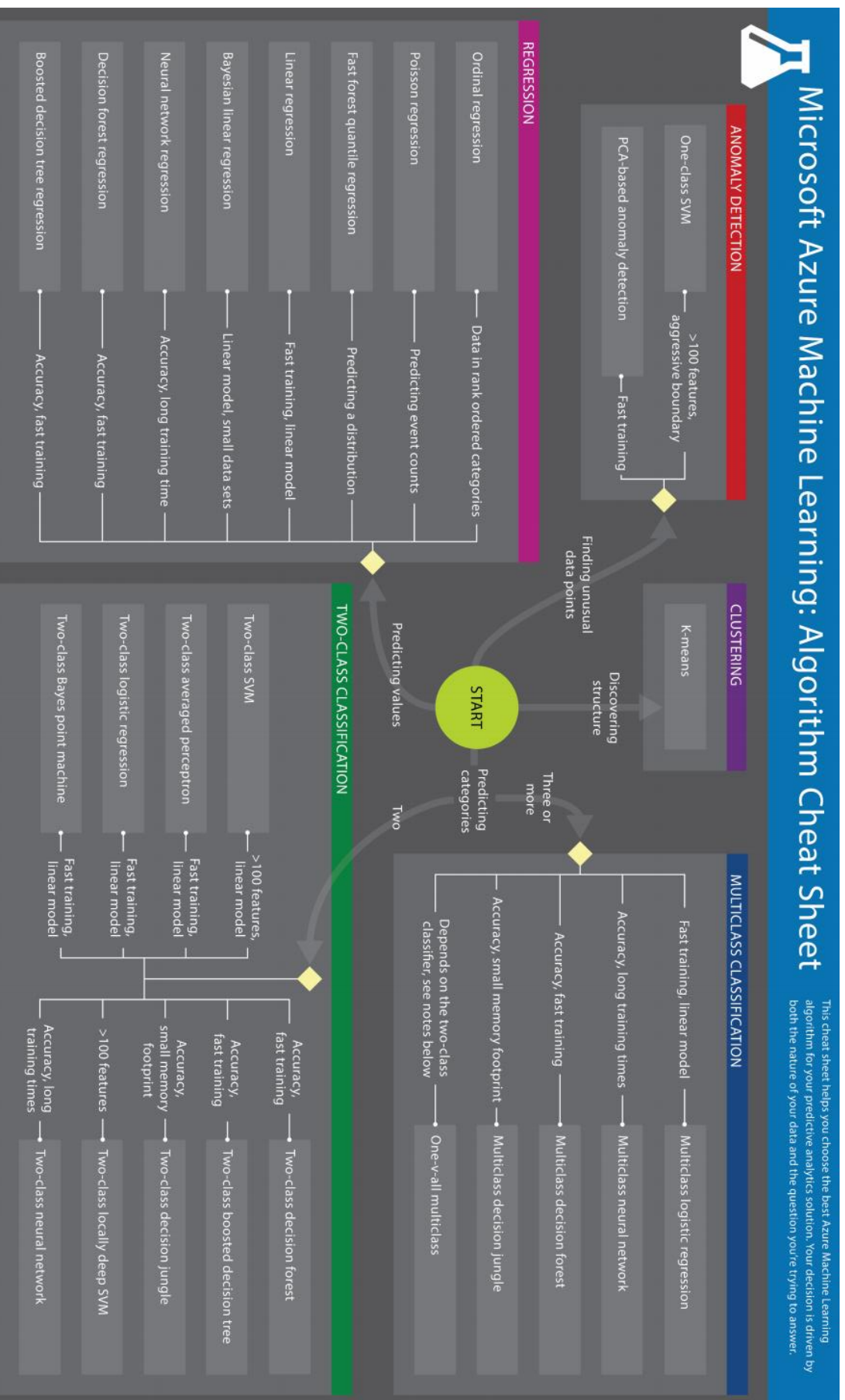
Tab. 3 Azure – algoritmy pre detekciu anomálií

- **Klasifikácia** - metóda strojového učenia, ktorá využíva údaje na určenie kategórie, typu alebo triedy skúmaných údajov. Klasifikácia sa môže využiť na:
 - Klasifikáciu emailov ako spam, nevyžiadané alebo dobré.
 - Určenie či pacientova vzorka je rakovinová.
 - Kategorizácia zákazníkov podľa ich sklonu reagovať na predajnú kampaň.
 - Identifikácia sentimentu ako pozitívny alebo negatívny.

Klasifikačné úlohy sa často organizujú podľa toho, či je klasifikácia binárna (A alebo B) alebo multiclass (možnosť predpovedania viacerých kategórií použitím jediného modelu) [6]. Nasledujúca tabuľka 4 zobrazuje algoritmy dostupné pre klasifikáciu.

Algorithm	Accuracy	Training time	Linearity	Parameters	Notes
Two-class classification					
logistic regression		•	•	5	
decision forest	•	○		6	
decision jungle	•	○		6	Low memory footprint
boosted decision tree	•	○		6	Large memory footprint
neural network	•			9	Additional customization is possible
averaged perceptron	○	○	•	4	
support vector machine		○	•	5	Good for large feature sets
locally deep support vector machine	○			8	Good for large feature sets
Bayes' point machine		○	•	3	
Multi-class classification					
logistic regression		•	•	5	
decision forest	•	○		6	
decision jungle	•	○		6	Low memory footprint
neural network	•			9	Additional customization is possible
one-v-all	-	-	-	-	See properties of the two-class method selected
Algorithm properties:					
• - shows excellent accuracy, fast training times, and the use of linearity					
○ - shows good accuracy and moderate training times					

Tab. 4 Azure – algoritmy pre klasifikáciu



Obr. 3 Azure - rozdelenie algoritmov pre strojové učenie

2.2 Trénovanie

Trénovanie je proces analýzy vstupných dát využitím preddefinovaného modelu. Z tejto analýzy sa model naučí vzory a uloží ich vo forme vytrénovaného modelu. Dôležité je správne zvoliť model na trénovanie, napr. pre detekciu anomálií by sa mal využiť model „Train Anomaly Detection Model“.

2.3 Skórovanie

Skórovanie v strojom učení sa používa na označenie procesu získavania nových hodnôt pomocou vytrénovaného modelu a nových vstupných dát.

2.4 Hodnotenie

Posúdenie vhodnosti natrénovaného modelu, ktoré sa vykoná po ukončení tréningu, pričom dôraz sa kladie na presnosť predpovedí, ktoré vieme z neho získať.

3 AWS – strojové učenie

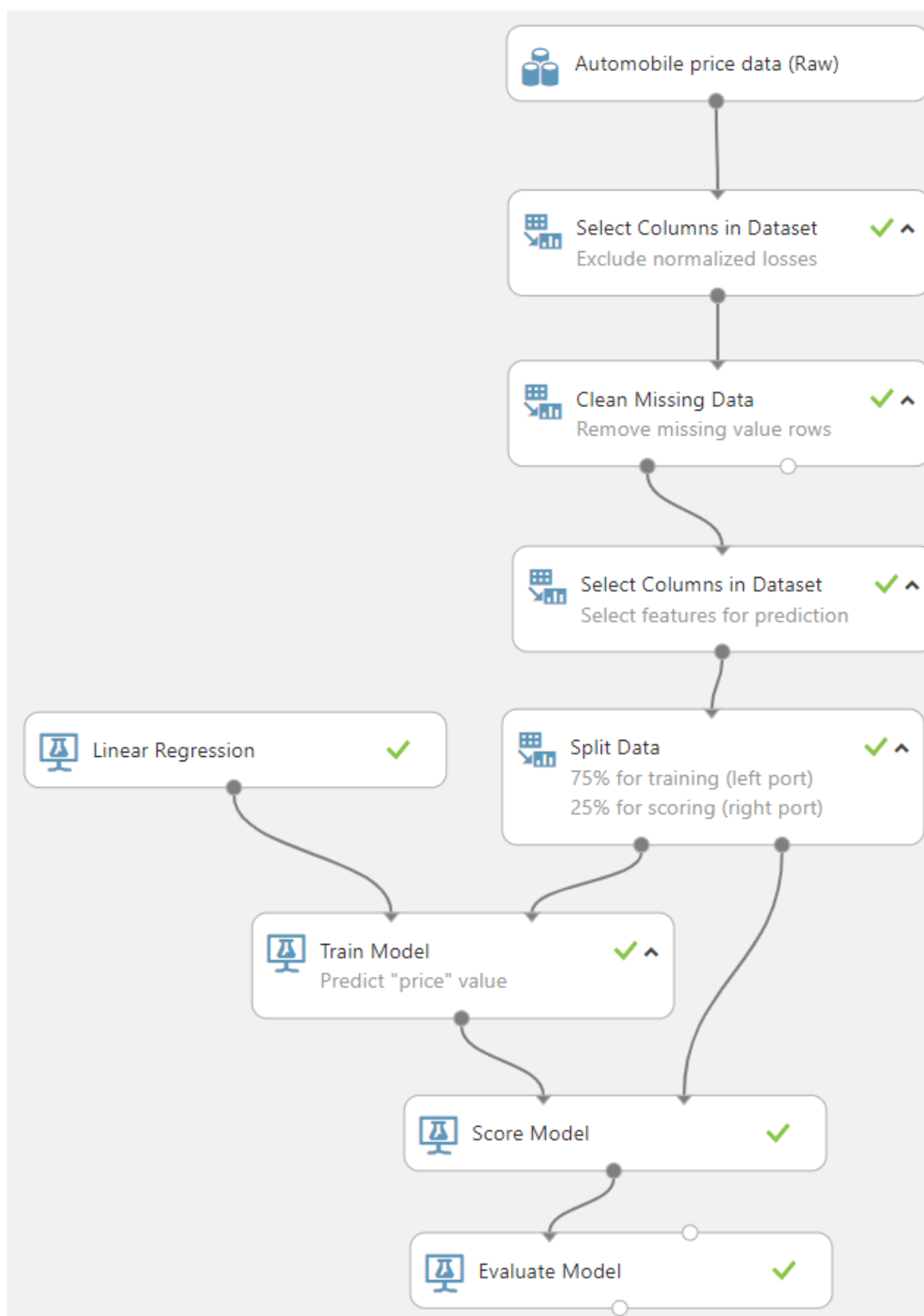
Amazon Web Services (AWS) je dcérska spoločnosť Amazon-u, ktorá poskytuje cloud computing-ové platformy na požiadanie jednotlivcom, spoločnostiam a vládam na základe predplatného.

AWS ponúka taktiež algoritmy strojové učenia, avšak ich ponuka sa skladá z troch modelov [7]:

- **Binárny klasifikačný model** - ML modely binárnych klasifikačných problémov predpovedajú binárny výsledok (A alebo B). Na tréning binárnych klasifikačných modelov Amazon ML používa algoritmus učenia v priemysle známy ako logistická regresia. Príklady binárnych klasifikačných problémov:
 - "Je tento e-mail spam alebo nie je spam?"
 - "Zakúpi zákazník tento produkt?"
 - "Je tento produkt kniha alebo hospodárske zviera?"
 - "Je táto recenzia napísaná zákazníkom alebo robotom?"
- **Multiclass klasifikačný model** - Umožňuje vytvárať predpovede pre viaceré triedy. Pre školenie multiclass modelov Amazon ML používa algoritmus učenia v priemysle známy ako multinomiálna logistická regresia. Príklady problémov s viacerými triedami:
 - "Je tento produkt knihou, filmom alebo oblečením?"
 - "Je tento film romantickou komédiou, dokumentárnym filmom alebo thrillerom?"
 - "Ktorá kategória produktov je pre tohto zákazníka najzaujímavejšia?"
- **Regresný model** - Predpovedajú číselnú hodnotu. Pre tréningové regresné modely Amazon ML používa algoritmus učenia v priemysle známy ako lineárna regresia. Príklady problémov s regresiami:
 - "Aká bude zajtra teplota v Bratislave?"
 - "Koľko kusov tohto produktu sa predá ?"
 - "Za akú cenu sa predá tento dom ?"

4 Predikcia ceny auta – príklad strojového učenia v Azure

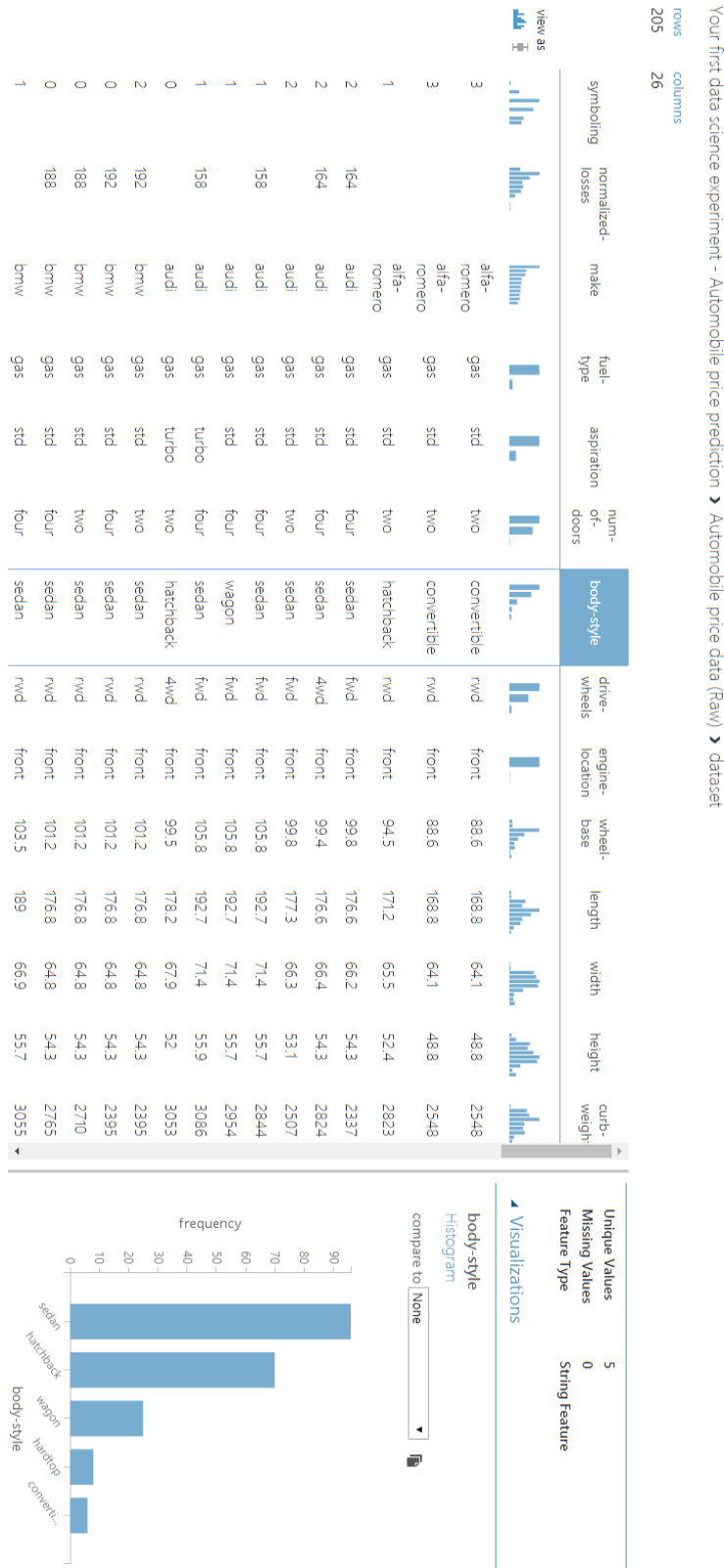
Nasledujúca kapitola obsahuje popis tvorby modelu strojového učenia pre predikciu ceny auta. Na obrázku 4 je možné vidieť celý model návrhu, ktorý je možné nasadiť ako webovú aplikáciu a získať predpokladanú cenu auta, na základe vložených informácií od užívateľa.



Obr. 4 Príklad - model strojového učenia pre predikciu ceny auta

Kompletný model predikcia ceny auta sa skladá z nasledujúcich celkov:

- **Automobile price data (Raw)** - obsahuje údaje o jednotlivých automobiloch vrátane informácií o značke, modeli, technických špecifikáciách a cene (Obr. 5).



Obr. 5 Príklad - údaje o jednotlivých automobiloch

- **Select Columns in Dataset** – V nasledujúcom bloku vyberieme údaje, ktoré chceme použiť pri trénovaní modelu. Nakoľko stĺpec „normalized-losses“ obsahuje veľa riadkov s chýbajúcimi údajmi, tak bude vynechaný.
- **Clean Missing Data** – Odstránenie riadkov s chýbajúcimi údajmi.
- **Select Columns in Dataset** – V nasledujúcom kroku je nutné zvoliť údaje o ktoré máme záujem, pričom každý riadok predstavuje jeden automobil a každý stĺpec je vlastnosťou tohto automobilu. Z celkovej počtu 26 stĺpcov/vlastností sme zvolili 8 (Obr. 6).

Select columns

BY NAME
WITH RULES

☐ Allow duplicates and preserve column order in selection

Begin With
ALL COLUMNS NO COLUMNS

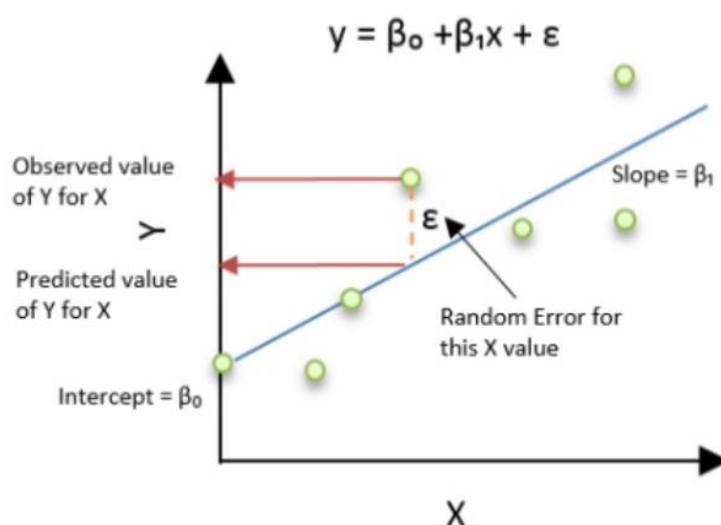
Include column names

make × body-style × wheel-base × engine-size ×
horsepower × peak-rpm × highway-mpg ×
price ×

+ -

Obr. 6 Príklad - zvolené vlastnosti automobilu

- **Linear Regression** - Pretože sa predpovedá cena, čo je číslo, tak sa využíva regresný algoritmus. Tento príklad používa jednoduchý model lineárnej regresie (Obr. 7).



Obr. 7 Príklad grafu lineárnej regresie

- **Split Data** – Rozdelenie údajov do dvoch skupín. Prvá skupina – 75% údajov sa používa na natrénovanie modelu a zvyšných 25% sa použije na otestovanie natrénovaného modelu.
- **Train Model** – V danom bloku je nutné zvoliť vlastnosť, ktoré chceme predikovať. V našom prípade je to cena. Model skenuje údaje a hľadá korelácie medzi vlastnosťami automobilu a jeho cenou.
- **Score Model** – Blok, v ktorom sa využije natrénovaný model a zvyšných 25% údajov, pričom jeho výstupom sú predpokladané/predikované ceny automobilov a známe ceny z testovacích údajov (Obr. 8).

Your first data science experiment - Automobile price prediction > Score Model > Scored dataset

rows 48 columns 9

view as  

make	body-style	wheel-base	engine-size	horsepower	peak-rpm	highway-mpg	price	Scored Labels
subaru	sedan	97	108	111	4800	29	11259	10286.204819
mitsubishi	hatchback	93.7	92	68	5500	38	6669	5446.847864
dodge	hatchback	93.7	90	68	5500	38	6229	6344.800711
honda	hatchback	86.6	92	76	6000	38	6855	5528.302953
alfa-romero	convertible	88.6	130	111	5000	27	16500	13498.476233
volvo	wagon	104.3	141	114	5400	28	16515	16097.608038

Známa hodnota (price) ↓ Predikovaná hodnota (Scored Labels)

Obr. 8 Príklad - skórovanie - známa hodnota vs predikovaná hodnota

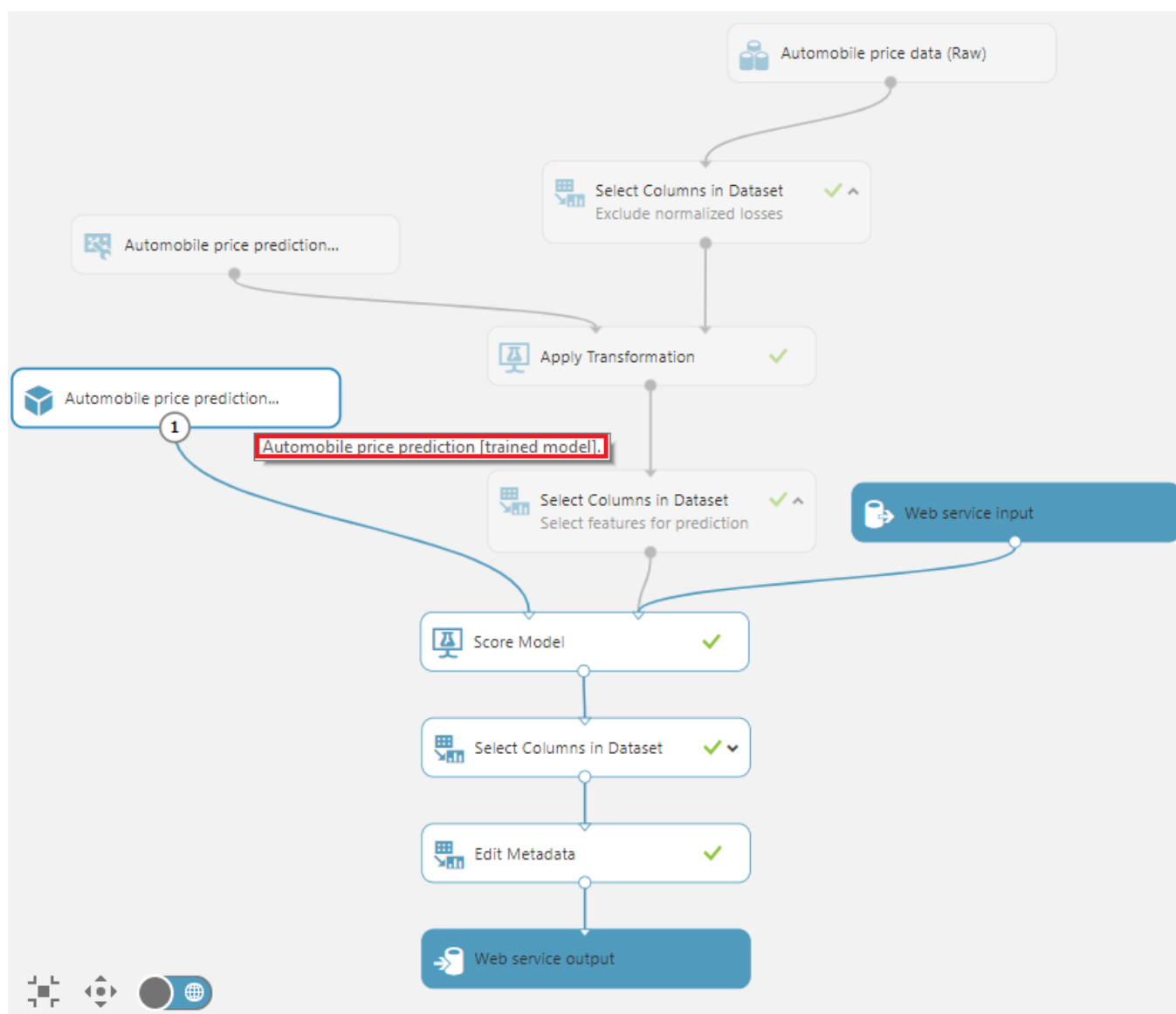
- **Evaluate Model** – Posledným krok je vyhodnotenie natrénovaného modelu (Obr. 9). Medzi zaujímavé parametre patrí MAE, ktorý predstavuje rozdiel medzi predpokladanou hodnotou a skutočnou hodnotou. Ďalším zaujímavým parametrom je napríklad Coefficient of Determination, ktorý predstavuje štatistickú metriku indikujúcu, ako dobre model opisuje zadané údaje.

Metrics

Mean Absolute Error	1656.147651
Root Mean Squared Error	2456.983209
Relative Absolute Error	0.276606
Relative Squared Error	0.089608
Coefficient of Determination	0.910392

Obr. 9 Príklad - Vyhodnotenie modelu

Natrénovaný model je ďalej možné využiť aj pri tvorbe webového servisu (Obr. 10), kde sa využije ako vstup do bloku „Score Model“ spolu s parametrami, ktoré zadá užívateľ, t. j. blok „Web service input“. Následne si zdefinujeme parametre, ktoré chceme zobrať ako odpoveď užívateľovi (blok „Select Columns in Dataset“). Nakoľko názov predikovanej ceny automobilu je vyjadrený ako „Score Labels“ (Obr. 8), tak vďaka bloku „Edit Metadata“ môžeme jeho názov zmeniť na „Predicted price of car“ (Obr. 11). Model webového servisu obsahuje aj pôvodný dátový set, nakoľko daný set slúži ako vzor pre štruktúru dát, ktoré model očakáva.



Obr. 10 Príklad – Webový servis s natrénovaným modelom

Štúdio pre strojové učenie v Azure ponúka možnosť si aj otestovať vytvorený webový servis. Užívateľ môže zadať parametre, ktoré servis očakáva a dostane predikovanú cenu automobilu (11). Nakoľko cena je hodnota, ktorú chceme zistiť, tak sa vyplní hodnotou 0.

input1

make

BMW

body-style

Sedan

wheel-base

88.6

engine-size

130

horsepower

111

peak-rpm

5000

highway-mpg

27

price

0

output1

make

BMW

body-style

Sedan

wheel-base

88.6

engine-size

130

horsepower

111

peak-rpm

5000

highway-mpg

27

Predicted price of car

17834.8794765581

Test Request-Response

Obr. 11 Príklad – Možnosť otestovať webový servis predikcie ceny auta

Záver

Cieľom umelej inteligencie je vytvoriť zariadenia, ktoré dokážu myslieť a správať sa tak, ako to robia ľudia. Na dosiahnutie tohto cieľu môžu byť použité rôzne technológie a jednou z nich je aj strojové učenie.

V súčasnej dobe veľké spoločnosti ako Google využívajú strojové učenie, aby sa učili zo získaných údajov a poskytli užívateľom lepšie výsledky vyhľadávania, a tým redukovali čas používateľa, ktorý musí stráviť vyhľadávaním požadovaných informácií.

Strojové učenie od Azure ponúka väčšie možnosti výberu algoritmov, ktoré je možné využiť pre natrénovanie modelu. Po nasedení webového servisu Azure automaticky vygeneruje kód v rôznych programovacích jazykoch, ktorý je možné využiť vo svojej aplikácii. Kód stačí jednoducho len prekopírovať. Výhodou riešenia od Azure je aj možnosť zdarma si vyskúšať ML bez nutnosti registrácie, t. j. nie je nutné vytvoriť účet, kde sa vyžaduje zadanie kreditnej karty. Dokumentácia k strojovému učeniu od Azure je kvalitná a poskytuje dostatočné informácie pre prácu s ním.

Smer pre budúci výskum a štúdium vidím v možnosti zadefinovania úloh v priemysle, ktoré by bolo možné riešiť pomocou modelu detekcie anomálií. Rozšírenie práce taktiež vidím aj v možnosti využitia nasadeného webového servisu v reálnej aplikácii.

Zdroje

- [1] Umelá inteligencia, strojové učenie, spracovanie prirodzeného jazyka - aký je rozdiel? [on-line]. [cit: 2018-28-10]. Dostupné na: <<https://sk.electronics-council.com/artificial-intelligence-16836>>
- [2] What is Azure? [on-line]. [cit: 2018-30-10]. Dostupné na: <<https://azure.microsoft.com/en-us/overview/what-is-azure> >
- [3] Machine Learning - Initialize Model - Regression. [on-line]. [cit: 2018-31-10]. Dostupné na: <<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-initialize-model-regression>>
- [4] Machine Learning - Initialize Model - Clustering. [on-line]. [cit: 2018-31-10]. Dostupné na: <<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-initialize-model-clustering>>
- [5] Anomaly Detection. [on-line]. [cit: 2018-1-10]. Dostupné na: <<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/anomaly-detection>>
- [6] Machine Learning - Initialize Model. [on-line]. [cit: 2018-2-10]. Dostupné na: <<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-initialize-model>>
- [7] Types of ML Models. [on-line]. [cit: 2018-4-10]. Dostupné na: <<https://docs.aws.amazon.com/machine-learning/latest/dg/types-of-ml-models.html#binary-classification-model>>