Predpovedanie návštevnosti obchodu z historických dát

Filip Matušák

Abstrakt

Cieľom tohto projektu bolo pokúsiť sa nájsť model, ktorý by čo najlepšie vedel predpovedať návštevnosť obchodu na základe historických dát, získaných z kamier nad vchodmi. Tieto kamery posielajú informácie o vstupe alebo výstupe z obchodu v reálnom čase. Hľadaný model by mal neskôr poslúžiť pri detekcií výpadkov kamier alebo iných anomálií.

Na agregovaných dátach z kamery sme sa pokúsili natrénovať a nájsť čo najlepšie parametre pre lineárnu regresiu aj náhodné lesy. Oba modely boli dosť úspešné. Pri lineárnej regresii bolo dosiahnuté skóre 0,83 a pri náhodných lesoch až 0,93.

Projekt je umiestnený na https://github.com/filipmatusak/projektMl2018

Úvod do problematiky

Pri real-time analytických softvéroch správania zákazníkov v kamenných obchodoch je dôležité mať korektné online dáta. Kedykoľvek však môže nastať výpadok rôzneho charakteru, kedy dáta nedotekajú alebo sú niečím silno ovplyvnené. Zachytávanie takýchto udalostí je dobré, čo najviac automatizovať. Jednou z možností detekcie výpadkov je porovnať aktuálne alebo nedávne dáta z kamery oproti očakávanému alebo predpovedanému stavu. Tu nám vie pomôcť strojové učenie. Preto sme hľadali čo najlepší regresný model, ktorý by predpovedal počet návštev v daný deň alebo hodinu.

Vstupné dáta

Zdroj

Spracovávali sme dáta z kamery, umiestnenej v jednom nemenovanom obchode na Obchodnej ulici v Bratislave. Surové dáta z kamery obsahujú veľa informácií, ktoré pri analýze dát nepotrebujeme. Napríklad rôzne informácie o prevádzke a kamere odkiaľ pochádzajú. Používame dáta od júna 2016 po december 2017. Počas zberu dát boli už historicky zaznamenané nejaké výpadky. Pravdepodobne by pomohli aj informácie o akciách v obchode, výpredajoch alebo počasí. Také dáta však nemáme k dispozícii.

Príprava dát

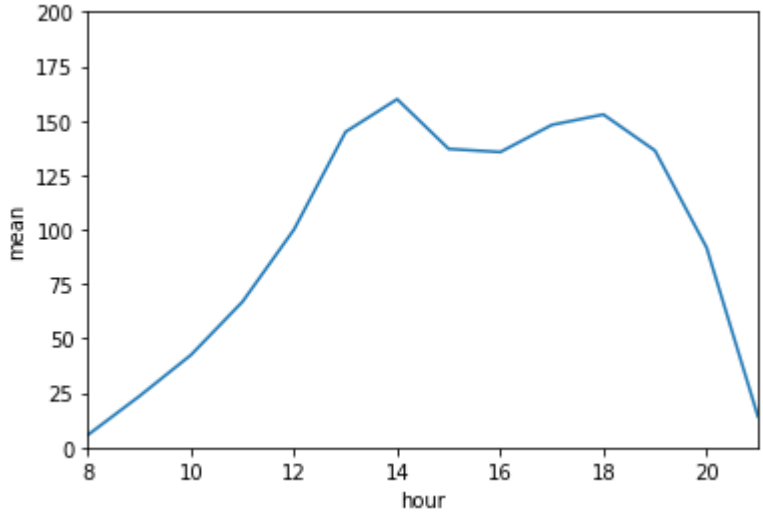
Kvôli možnému skresleniu sme odstránili z dát záznamy z dní, kedy boli výpadky. Nakoniec sme tieto dáta agregovali po hodinách a odstránili sme z nich nepotrebné informácie. Zostali nám záznamy s týmito údajmi:

- rok
- mesiac v roku
- deň v mesiaci
- hodina v dni
- deň v týždni
- počet návštev

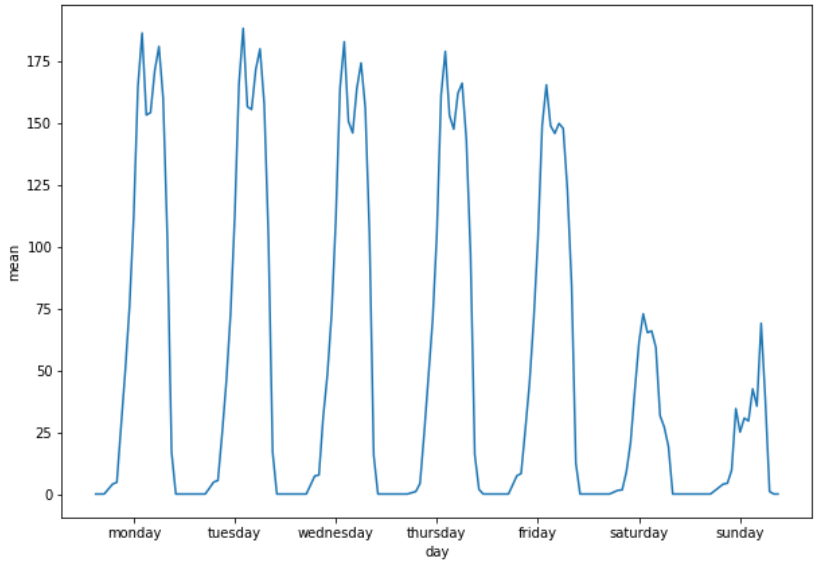
Charakteristika dát

Hodnoty všetkých atribútov sú prirodzené čísla. Každý z nich môže mať prirodzene vplyv na počet návštev.

Na nasledujúcom grafe je vidieť, ako vyzerá priemerný deň v obchode.



Na ďalšom grafe je znázornený priemerný týždeň.



Implementácia

Prvá časť, získanie dát, bola naimplementovaná v jazyku Scala s použitím frameworku Spark. V tejto časti sa načítajú surové dáta z databázy Cassandra. Následne sa z dát vyberie spomínaný časový interval a vyfiltruje sa od výpadkov. Výsledné dáta sa zapíšu do súboru.

Na implementáciu trénovania modelov a reporty sme použili programovací jazyk Python spolu s knižnicami sklearn, matplotlib, numpy, numpy_indexed. Výsledné vyhodnotenie modelu sme robili pomocou funkcie ‘score‘, ktorá pre daný model vráti koeficient determinácie R^2. Najlepšie skóre môže byť 1. Heuristika, ktorá vracia vždy len očakávanú cieľovú hodnotu má skóre 0.

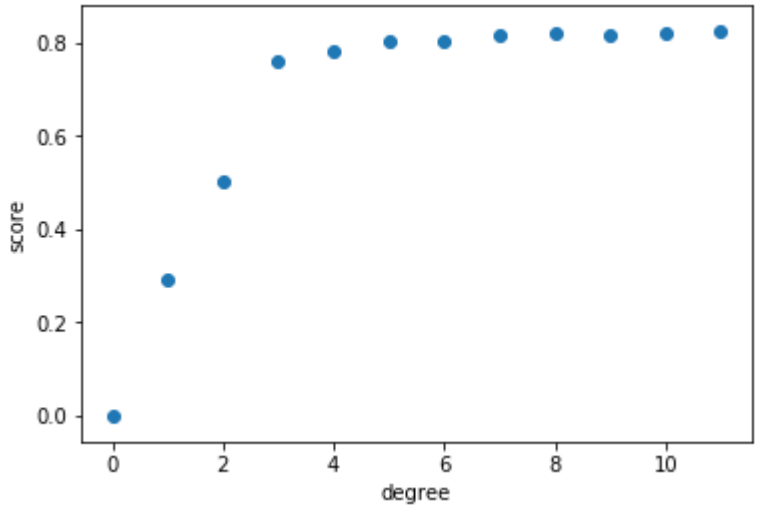
Pri trénovaní sme používali rozdelenie dát na 80% trénovacích dát a 20% testovacích, pri klasickom testovaní (fit, score), ale aj pri cross validation.

Lineárna regresia

Použili sme implementáciu LinearRegression zo sklearn.

Lineárna regresia sa dá použiť skoro na každý regresný problém. Treba však mať vhodné atribúty, alebo kombinácie atribútov. Preto sme museli vyskúšať rôzne polynomiálne rozvoje atribútov, aby sme našli správne parametre lineárnej regresie. Na nasledujúcom grafe je znázornená testovacia chyba pre lineárnu regresiu s

polynomiálnym rozvojom atribútov od 1 po 10. Vyššie rozvoje už majú priveľké časové a pamäťové nároky.

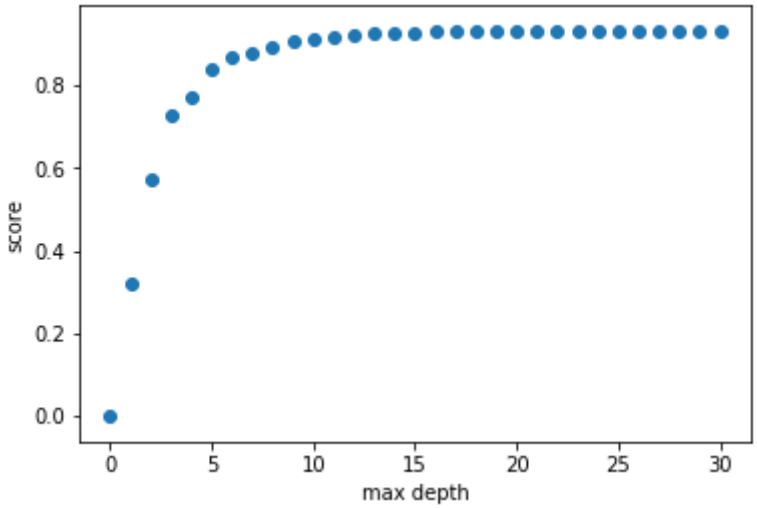


Od rozvoja do tretieho stupňa sa chyba pomaly stabilizuje. Najlepšie výsledky dosahuje pri rozvoji do ôsmeho stupňa. Vtedy dosahuje lineárna regresia skóre 0,822. Podľa cross validácie je to až 0,83.

Samozrejme najdôležitejším atribútom na predpoveď je hodina v dni.

Náhodné lesy

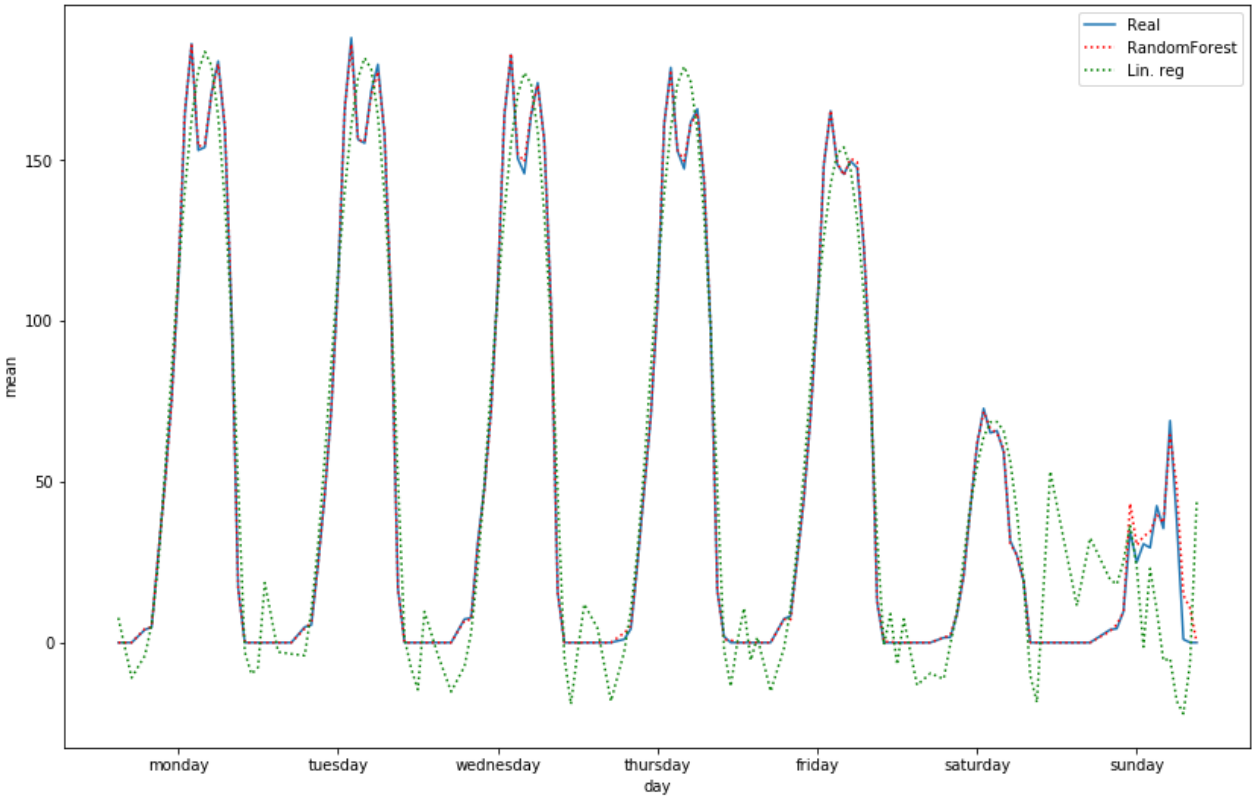
Druhým testovaným modelom boli náhodné lesy, konkrétne RandomForestRegressor zo sklearn. Existuje veľa parametrov, ktoré ovplyvňujú úspešnosť tohto modelu. Pre skoro všetky sme použili pôvodné nastavenia z knižničnej implementácie lesu. Nastavili sme počet stromov na 20. A jediným hľadaným parametrom bol pre nás maximálna hĺbka stromu. Natrénovali a otestovali sme RandomForestRegressor pre hodnoty maximálnej hĺbky od 1 do 30. Pre každý z nich sme vypočítali skóre. Nasledujúci graf zobrazuje tieto výsledky.



Trénovanie lesov trvalo oveľa menej ako lineárna regresia pre veľký rozvoj. Od maximálnej hĺbky 10 sa chyba stabilizuje. Najlepšie skóre dosahoval RandomForestRegressor pre maximálnu hĺbku stromu 16. Scóre v tomto prípade bolo 0,93.

Výsledky

Z vypočítaného scóre oboch modeloch je vidieť, že oba sú úspešné v predpovedaní pre dané dáta. Dôvodom môže byť, že dáta z tohto obchodu sa správajú slušne, závisia hlavne od času a nie sú silno ovplyvňované napríklad počasím. Pre lineárnu regresiu bolo skóre 0,83 a pre náhodne lesy 0,93. Heuristika, ktorá vracia priemernú hodnotu má skóre 0. Lineárna regresia počíta výslednú hodnotu ako lineárnu kombináciu atribútov. Výhodou rozhodovacieho stromu je, že ak sa v nejakom vrchole rozvetví podľa niektorého artibútu, potom podstrom tohto vrcholu je model pre dáta s konkrétnym ohraničením hodnôt atribútov. Preto napríklad RandomForestRegression vie lepšie predpovedať návštevnosť, ak si dáta rozdelí podľa toho, či je víkend alebo pracovný deň. Jasné rozdiely v oboch modeloch je vidieť na ďalšom grafe, na ktorom sú zobrazené dáta pre priemerný týždeň predpovedané jednotlivými modelmi a reálne dáta.



Môžeme vidieť, že aj skóre 0,83 nemusí byť dosť.

Záver

Na dátach o návštevnosti obchodu sme sa pokúsili natrénovať lineárnu regresiu aj náhodné lesy. Našli sme parametre, pri ktorých mali modely najlepšie skóre na daných dátach. Nakoniec sa ako lepšie riešenie ukázali náhodné lesy pred lineárnou regresiou. Ukázalo sa to nie len v skóre (0,93 oproti 0,83) ale aj na grafe. Pre ešte lepšiu predpoveď by nám pomohli dodatočné dáta napríklad o počasí.

Myslíme, že projekt bol úspešný. Podarilo sa nám nájsť dosť presný model na predpoveď návštevnosti daného obchodu. Tento model môžeme použiť pri hľadaní anomálií v zbere dát z tohto obchodu.

Použitá literatúra

- Dokumentácia numpy - https://docs.scipy.org/doc/
- Dokumentácia sklearn - http://scikit-learn.org/stable/documentation.html
- Dokumentácia matplotlib - https://matplotlib.org/contents.html