Primena Mašinskog Učenja za Prognoziranje Prodaje u Vremenskim Serijama

Luka Mihajlović Vojislav Stevanović Andrija Ilić

2020/0136 2021/1070 2020/0236

Jun 2024

1 Uvod

U promenljivom svetu maloprodaje, precizno predviđanje količine prodaje je supermoć. Zamislite da znate šta, kada i koliko da skladištite kako biste savršeno zadovoljili potražnju kupaca. Ovde dolazi do predviđanje prodaje u kontekstu vremenskih serija.Vremenske serije u prodaji predstavljaju niz podataka o prodaji koji su uređeni hronološki, obično na dnevnom, nedeljnom, mesečnom ili godišnjem nivou. Analiza ovih podataka omogućava identifikovanje obrazaca, trendova i sezonalnosti, što je ključno za predviđanje buduće prodaje.

1.1 Opis problema

Cilj je predviđanje buduće prodaje u "Favorita", lancu prehrambenih prodavnica iz Ekvadora. Skup podataka obuhvata istorijske podatke o prodaji, zajedno sa informacijama o prodavnicama, tipovima proizvoda i promocijama.

1.2 Izazovi u rešavanju problema

Priroda vremenskih serija - podaci o prodaji često pokazuju trendove, sezonalnost i nepredvidive promene. Efikasno baratanje ovim faktorima, kao i hvatanje obrazaca ključno je za precizno predviđanje.

Više faktora - na prodaju utiču razni faktori, uključujući promocije, praznike i lokaciju prodavnice. Model treba da uzme u obzir što više delotvornih faktora radi boljeg predviđanja

Više izlaza - predviđanje prodaje za velik broj kategorija proizvoda unosi dodatnu složenost u poređenju sa predikcijama u slučaju singularnog izlaza.

1.3 Zašto je problem bitan?

Tačno predviđanje prodaje od suštinskog je značaja za maloprodajna preduzeća. Pomaže im u:

- Optimizaciji upravljanja zalihama: predviđanjem potražnje, mogu se izbegnuti preterane zalihe i sprečiti zalihe onih artikala koji su skloni kvarenju, ili ograničenog roka trajanja.

- Poboljšanju promotivnih strategija: razumevanje kako promocije utiču na prodaju omogućava bolje planiranje i raspodelu resursa unutar lanca prodaje

- Poboljšanju donošenja odluka: pouzdane prognoze informišu menadžment kako bi se mogle doneti strateške odluke o ponudi proizvoda, nivou osoblja, kao i samim cenama ponude.

1.4 Kome je bitan?

- Prodavcima - Oni stiču uvide u optimizaciju zaliha, promocija i celokupno poslovne strategije. - Potrošačima - Koriste prednosti poboljšane dostupnosti proizvoda i potencijalno nižih cena zasnovanih na tačnim predviđanjima. - Praktikantima mašinskog učenja - Pružena im je platforma za vežbanje i usavršavanje veština predviđanja na podacima koji su realni.

2 Opis podataka

Podaci su preuzeti sa sledećeg linka: Store Sales - Time Series Forecasting.

Skup podataka se sastoji od sledećih CSV fajlova koji opisuju prodaju u lancu prehrambenih prodavnica "Favorita" u Ekvadoru:

- train.csv i test.csv: Sadrže podatke o dnevnoj prodaji po prodavnicama i kategorijama proizvoda, uključujući informacije o promocijama. Cilj analize je predvideti vrednosti u koloni sales.

- stores.csv: Pruža dodatne informacije o prodavnicama, kao što su lokacija (grad i država), tip prodavnice i grupa sličnih prodavnica (klaster).

- oil.csv: Sadrži dnevne podatke o ceni nafte (West Texas Intermediate Crude Oil), što može biti relevantno za analizu uticaja ekonomskih faktora na prodaju (Ekvador je država koja dosta zavisi od nafte i njena ekonomska dobrobit je vrlo osetljiva na povećanja u ceni nafte).

Skup podataka ne sadrži nedostajuće vrednosti, osim u podacima vezanim za cenu nafte (dcoilwtico).

2.1 Vizuelizacija prodaje

2013: Prodaja je relativno niska i stabilna tokom većeg dela godine, sa porastom krajem godine.

2014: Prodaja počinje da raste, sa izraženijim sezonskim fluktuacijama.

2015: Prodaja nastavlja da raste, dostižući novi vrhunac krajem godine.

2016: Prodaja dostiže najviše vrednosti u posmatranom periodu, ali sa fluktuacijama tokom godine.

2017: Prodaja se stabilizuje na visokom nivou, sa manjim fluktuacijama nego u prethodnoj godini.

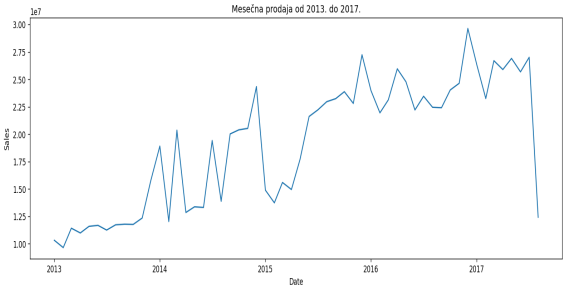

Figure 1: Mesečna prodaja u periodu od 2013. do 2017.

3 Priprema podataka

S obzirom na postojanje različitih .csv file-ova, za predikciju nam je potrebno da se podaci nalaze u jednom Data frame-u. S tim na umu, sprovedeno je merge-ovanje svih .csv file-ova, grupisanih po datumu i prodavnici, u jedan finalni data frame iz kog ćemo kasnije izvoditi atribute. Nakon što je merge urađen, pojavile su se nedostajuće vrednosti koje nisu prethodno bile prisutne. Te su nedostajuće vrednosti došle iz holiday.csv skupa podataka, kao rezultat nepostojanja praznika za svaki datum koji je postojao u originalnom skupu podataka. Kategoričke kolone sređene su putem kreiranja tzv. "dummy" varijabli. Kada su one sređene, iz kolone datuma izvedeni su atributi - day_of_week, day_of_year, day_of_month, month, quarter, year. Oni će pomoći u treniranju modela.

4 Treniranje i interpretacija rezultata

Nakon što je skup podataka podeljen (80% train, 20% test), napravljena su tri bazična modela skladna ovom problemu. Random Forest, Linearna regresija i XGBoost su trenirani u svrhe poređenja i evaluacije predviđanja. Za evaluaciju modela korišćene su sledeće metrike:

Koren iz srednjeg kvadratnog odstupanja:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$$

Srednje apsolutno odstupanje

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|$$

R^2 mera

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}$$

Prvobitno su trenirani sa default parametrima, i performanse su bile ovakve:

Table 1: Performanse modela za predviđanje prodaje

Model	RMSE	MAE	R^2
Random Forest	3222.18	1955.00	0.920
LRegression	5032.35	3770.98	0.805
XGBoost	3177.99	1875.61	0.922

Naposletku je izvršena optimizacija hiper parametara modela, i za to se koristila nasumična pretraga sa unakrsnom validacijom. Ova metoda je izabrana kao dobra alternativa Grid pretrazi, koja je relativno neefikasna jer prolazi kroz sve moguće kombinacije hiperparametara, dok nasumična pretraga funkcioniše putem raspodele i verovatnoća, i na određen postavljen broj iteracija istražuje opcije.

Nakon optimizacije hiperparametara, performanse su sledeće:

Table 2: Performanse modela za predviđanje prodaje

Model	RMSE	MAE	R^2
Random Forest	3286.119	1995.151	0.917
LRegression	5032.348	3770.979	0.805
XGBoost	3161.902	1954.744	0.923

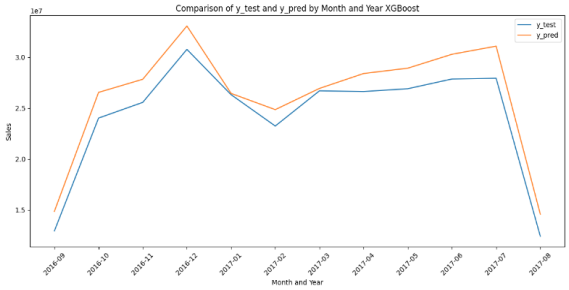

Figure 2: Poređenje između istinskih vrednosti i predviđanja za XGBoost

Uopšteno govoreći, svaki od modela je relativno dobar i po R^2 vidimo da sa vrlo prihvatljivom tačnošću izvode izlaznu promenljivu iz podataka, dok se XGBoost pokazuje kao najbolji, jer su i greške koje u proseku pravi najmanje.

Kao eksperiment, sprovedena je selekcija atributa i izmerena je delotvornost te selekcije na treniranje i evaluaciju modela. Korišćena je metoda varijanse, gde su isključeni podaci koji ne dodaju preterani varijabilitet skupu podataka, i korišćena je selekcija atributa putem feature importance-a. Ove metode imale su kao rezultat poboljšanja u performansama za XGBoost i Random Forest, kao što se može primetiti na sledećoj tabeli:

Table 3: Performanse modela za predviđanje prodaje

Model	RMSE	MAE	R^2
Random Forest	3264.626	1980.722	0.918
LRegression	5163.687	3924.267	0.795
XGBoost	2869.036	1794.925	0.937

Ovo se najverovatnije događa zbog prirode naših algoritama. Random Forest i XGBoost su algoritmi bazirani na stablima odlučivanja, a koriste bagging i boosting, respektivno. Algoritmi bazirani na stablima odlučivanja često imaju značajnih benefita od većeg broja atributa. U ovom slučaju je veliki deo atributa uskraćen, ipak to ima pozitivan uticaj na performanse modela, tako da se može relativno slobodno zaključiti da isključeni atributi nisu bili od naročite koristi tokom treniranja modela.

5 Zaključak

Nakon treniranja i optimizacije modela, treba ih staviti u produkciju. U ovu svrhu uzećemo XGBoost model jer je naposletku davao najbolje rezultate. Iako on nema mogućnost budućeg predviđanja, stream-ovanjem podataka u model možemo ostvariti real-time predikcije.

Ubuduće bi trebalo testirati još koji model, na primer modele usko skrojene za predviđanje u slučaju vremenskih serija. Modeli koji mogu biti korišćeni su: Prophet, ARIMA, SARIMAX ili neka vrsta rekurentne ili LSTM neuronske mreže. Takođe bi se mogla primeniti neka metoda smanjenja dimenzionalnosti podatka, konkretno analiza glavnih komponenti ili t-SNE.

Reference

[1] Jason Brownlee. *Random Forest for Time Series Forecasting*. `https://machinelearningmastery.com/random-forest-for-time-series-forecasting/`. Accessed: 2024-06-01. 2017.

[2] Jason Brownlee. *XGBoost for Regression*. `https://machinelearningmastery.com/xgboost-for-regression/`. Accessed: 2024-06-01. 2016.

[3] Dataquest. *Understanding Regression Error Metrics*. `https://www.dataquest.io/blog/understanding-regression-error-metrics/`. Accessed: 2024-06-01. 2018.

[4] Kaggle. *Store Sales - Time Series Forecasting*. `https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data`. Accessed: 2024-06-01.

[5] Mikio L. Lang **and** Barbara Lüdtke. "A model-free approach to anomaly detection for sequence data". **in***Machine Learning*: 109.6 (2020), **pages** 1047–1070. URL: `https://link.springer.com/article/10.1007/s10994-020-05910-7`.