Rossmann Store Sales: EDA i razvoj prediktivnih modela

Drobnjaković Tamara, Anđela Dimić, Dragica Draškić

1. Opis problema

Rossmann je lanac prodavnica sa hiljadama lokacija širom Evrope. Uprava kompanije želi da unapredi planiranje resursa i zaliha tako što će moći da predviđa dnevnu prodaju svake prodavnice.

Dostupni podaci obuhvataju istorijske podatke o prodaji, informacijama o promocijama, državnim praznicima, danima u nedelji, sezonalnosti, konkurenciji, tipu prodavnice, tipu asortimana…

Zadatak je da se na osnovu ovih karakteristika izgradi model mašinskog učenja koji predviđa da li će prodaja određenog dana biti veća ili manja od proseka za prethodnih 30 dana. Ovaj zadatak spada u problem klasifikacije, jer ćemo prodaju klasifikovati u jednu od dve moguće grupe: 1 (prodaja će biti veća od proseka) i 0 (prodaja će biti manja od proseka).

2. Opis i razumevanje podataka

Na raspolaganju nam stoje 3 dataset-a: train, test i stores. U *train* skupu podataka nalaze se informacije o poslovanju određene prodavnice određenog dana. Imamo uvid u to da li je radnja bila otvorena, koliko je mušterija došlo tog dana, kolika je vrednost prodaje, da li se sprovodila promocija itd. *Stores* skup podataka nam daje bliže informacije o konkretnoj prodavnici: kog je tipa, koji asortiman pruža, periode u kojima je sprovodila promociju kao i informacije o njenoj konkurenciji. *Test* skup nema ciljanu promenljivu Sales, sto znači da ne možemo oceniti performanse modela koji nad njegovim podacima vrše predviđanja. Dakle, taj dataset simulira realni svet.

Da bismo bolje razumeli ciljanu promenljivu Sales, mozemo predstaviti njene vrednosti kroz dve godine za dve nasumične prodavnice:

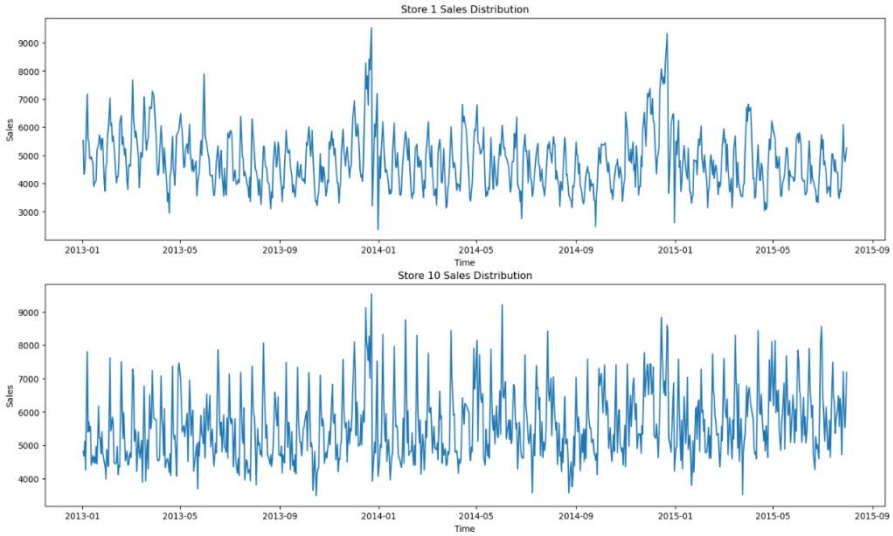


Image 1 – Grafik kretanje prodaje kroz vreme

Može se uočiti da:

- prodaja pokazuje izraženu sezonalnost i fluktuacije kroz vreme,
- postoje periodi naglog rasta i pada prodaje, što može biti posledica promotivnih aktivnosti, praznika ili drugih faktora,
- iako su obe prodavnice u istom lancu, njihovi obrasci prodaje se razlikuju, što ukazuje na uticaj lokalnih uslova (lokacija, konkurencija, tip kupaca, i dr.)

Eksploratorna analiza nam je omogućila da se upoznamo i sa ostalim promenljivama. Za numeričke kolone nas je interesovala raspodela vrednosti. Na osnovu histograma i boxplot-a možemo uočiti da i Sales i Customers kolone imaju asimetrične raspodele i puno outlier-a. Razlog tome jeste što je mnogo češća niska vrednost prodaje i mali broj kupaca, ali se ponekad jave i velike vrednosti ovih promenljivih. Kada su u pitanju kategoričke promenljive, zanima nas učestalost svake kategorije. Na countplot-ovima možemo primetiti da postoji značajno veći broj zapisa u kojima je radnja otvorena i u kojima nema promocije, a pojava bilo kog državnog ili školskog praznika je retka.

Da bismo uočili koliko su tip prodavnice i promocija važni za prodaju, kreirali smo sledeće dijagrame:

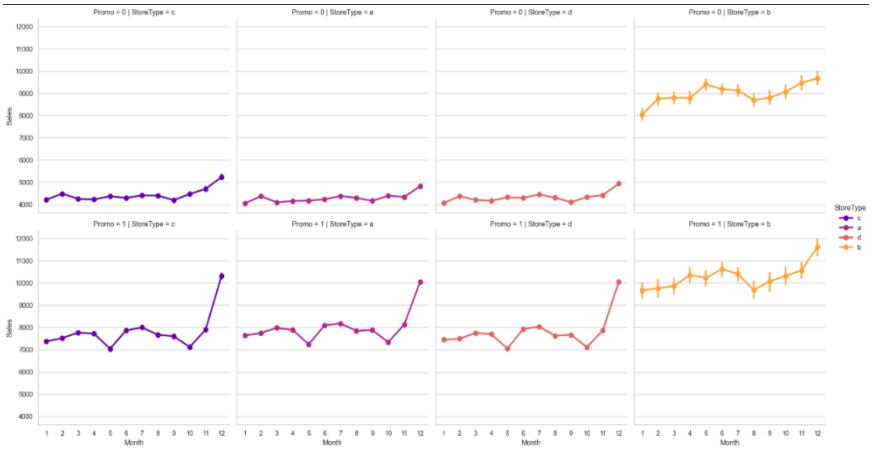
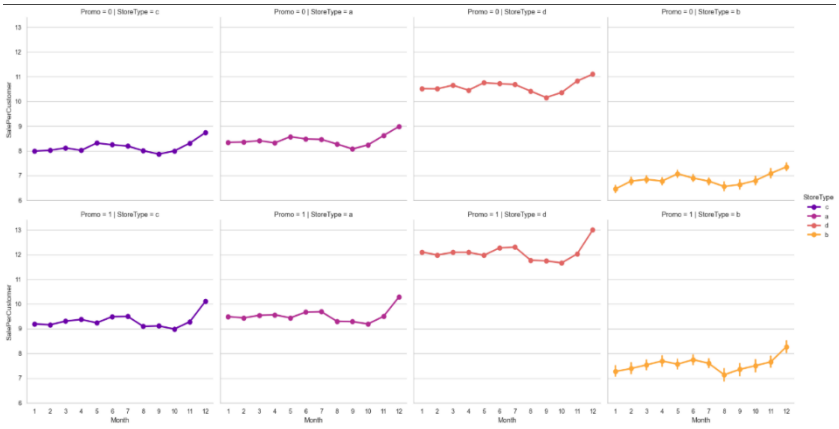


Image 2 - Grafik Kretanje prodaje u zavisnosti od tipa prodavnice i promocije

Zaključak:

- Promocija je izuzetno efikasna i predstavlja glavni faktor za povećanje prodaje u svim prodavnicama
- StoreType b je dominantan u smislu apsolutne prosečne prodaje
- Decembar je mesec gde vidimo skok prodaje u svim prodavnicama

Zatim smo posmatrali kako ova kombinacija atributa utiče na pomoćnu promenljivu SalesPerCustomer:



Iako su prethodni grafikoni pokazali da StoreType B ima najveću ukupnu prodaju, u stvarnosti on ima najnižu prodaju po kupcu (SalePerCustomer), što ukazuje na to da kupci kupuju male količine. StoreType D je, s druge strane, najprofitabilniji po kupcu, sa prosečnom potrošnjom od oko 12€ uz promociju.

Image 3 - Grafik Kretanje prodaje po kupcu u zavisnosti od tipa prodavnice i promocije

3. Pretprocesiranje podataka

3.1 Outliers

Numeričke promenljive Sales i Customers možemo optimizovati za modele previđanja tako što ćemo regulisati uoutlier-e. Ovo smo odlučili da uradimo putem ograničavanja (Capping). Umesto da u potpunosti uklonimo ekstremne vrednosti, mi bismo ih zamenili donjom, odnosno gornjom granicom za procenjivanje outlier-a. Međutim, ovakva transformacija nije neophodna jer se te kolone moraju izbaciti iz skupa podataka pre treniranja. One direktno ukazuju na rezultat našeg predviđanja, što predstavlja curenje podataka.

3.2 Nedostajuće vrednosti

Povezali smo train i test dataset pojedinačno sa store datasetom, koristeći left merge.

Primetili smo da se u kolonami PromoInterval nalazi veliki broj NaN vrednosti. Daljom analizom smo saznali da je to posledica nepostojanja dugoročnog promotivnog programa, pa ćemo te vrednosti zameniti sa ‘No promo’. NaN vrednosti u kolonama Promo2SinceYear i Promo2SinceWeek zamnićemo vrednošću 0.

Isti problem se javlja i sa kolonama CompetitionOpenSinceYear i CompetitionOpenSinceMonth. Utvrdili smo da su ovo zaista nedostajuće vrednosti, jer konkurencija postoji. Nedostajuce vrednosti atributa CompetitionOpenSinceYear ćemo aproksimirati medijanom, a za nominalni atribut CompetitionOpenSinceMonth koristicemo mesec koji se najčešće pojavljuje.

3.3 Generisanje target kolone i izdvajanje podskupa podataka

Cilj nam je da predvidimo da li će prodaja u određenoj radnji narednog dana biti veća ili manja od proseka u prethodnih 30 dana.

Za kreiranje takvog modela koristićemo podskup naseg dataset-a. U obzir ćemo uzeti 10 prodavnica za koje postoji najviše redova u dataset-u. Takođe, posmatraćemo samo podatke iz poslednje dve godine.

Kolonu Date transformisali smo u kolone Day, Month, Year i IsWeekend. U koloni StateHoliday vrednosti 0, a, b i c smo pretvorili u none, public_holiday, Easter_holiday i Christmas, respektivno. Konačno, kolonu Assortment smo umesto u obliku a, b i c zapisali kao basic, extra i extended.

3.4 Kodiranje i skaliranje

Kategorički atributi kodirani su One-Hot metodom kako bi ih modeli mogli koristiti kao numeričke ulaze. Numerički atributi imaju različite skale ali korišćeni algoritmi nisu osetljivi na različite opsege pa nije neophono izvršiti normalizaciju.

3.5 Selekcija atributa

Značajnost numeričkih atributa ispitivali smo u okviru korelacione matrice, a za kategoričke atribute smo koristili vizualni prikaz distribucije u odnosu na izlazni atribut. Primećujemo da se vrednosti Sales atributa mogu varirati manje ili vise u odnosu na većinu kategoričkih ordinalnih atributa, te da skoro svi atributi imaju barem neki uticaj na prodaju. Najvažniji uvidi su: Promocije (Promo=1) drastično povećavaju prodaju. Najveća prodaja je prvog dana u nedelji (Dan 1), dok je najmanja vikendom (Dani 6 i 7). Tip prodavnice 'a' ostvaruje znatno veći prosek prodaje od tipa 'd'. Prodaja je najveća u decembru (Mesec 12), a najviša prosečna prodaja tokom promotivnih intervala se ostvaruje u februaru, maju, avgustu i novembru. Numerički atributi imaju manji uticaj ali su i oni uključeni u kreiranje modela kako bismo otkrili sve potencijalne zakonitosti. Atribut Pormo2SinceWeek je izostavljen usled snažne korelacije sa drugim atributima. Atribute ID i Store smo uklonili zato što oni predstavljaju jedinstvene identifikatore. Sales i 30_a su atributi koje smo koristili za kreiranje izlaznog atributa pa cemo ih zbog toga ukloniti. Atribut Customers se ne nalazi u test setu i direktno je korelisan sa izlaznim atributom pa ga je potrebno izostaviti prilikom modelovanja. Uklonili smo i Date varijablu jer smo nju iskoristili za kreiranje novih atributa.

3.6 Balansiranje klasa

Odabrali smo nad-uzorkovanje kao tehniku kojom se ne gubi deo informacija sobzirom da smo ograničili tj. već smanjili naš skup podataka. SMOTE generiše nove sintetičke uzorke i kao takav predstavlja bolju opciju od Random Oversamplinga koji duplira već postojeće uzorke i time povećava šansu od overfitinga.

4. Odabrani modeli i podešavanje hiperparametara

U okviru modelovanja primenjena su tri klasifikaciona algoritma: **Decision Tree, Random Forest** i **AdaBoost**. Svi modeli su implementirani u okviru jedinstvenog *pipeline*-a koji obuhvata unapred definisan postupak obrade podataka. Pre obuke modela sprovedena je imputacija nedostajućih vrednosti, kodiranje kategorijskih promenljivih pomoću *One-Hot Encoding*-a i balansiranje klasa metodom **SMOTE**, čime je obezbeđeno da svi modeli rade nad istim, dosledno pripremljenim podacima.

Decision Tree model je izabran kao jednostavan i interpretabilan algoritam koji omogućava brzo testiranje odnosa između atributa i ciljne promenljive. Za njega su podešavani hiperparametri kao što su kriterijum podele (gini ili entropy), maksimalna dubina stabla, minimalan broj uzoraka potrebnih za podelu, minimalan broj opservacija u listovima i broj atributa koji se razmatra pri svakoj podeli.
Random Forest je korišćen kao proširenje stabla odlučivanja, sa ciljem da se smanji varijansa i poveća stabilnost modela. Tokom pretrage optimizovani su hiperparametri poput broja stabala u šumi i ostali hiperparametri koji su prisutni u klasičnom stablu odlučivanja. Ovaj model je posebno pogodan za kompleksne skupove podataka sa većim brojem atributa.
AdaBoost je korišćen kao predstavnik ansambl metoda koje kombinuju više slabih klasifikatora kako bi se postigla veća tačnost. U procesu podešavanja hiperparametara menjan je broj slabih klasifikatora i *learning rate*, koji kontroliše doprinos svakog klasifikatora konačnoj predikciji.

Za sve modele sprovedena je optimizacija hiperparametara pomoću **GridSearchCV** procedure uz **Stratified K-Fold (k=5)** unakrsnu validaciju, a kao kriterijum za izbor najboljih parametara korišćen je **F1-score**, jer uravnoteženo meri preciznost i odziv modela. Najbolje konfiguracije hiperparametara su zatim upotrebljene za evaluaciju modela na izdvojenom test skupu.

5. Evaluacija rešenja

5.1 Poređenje rezultata

Skup podataka je, zahvaljujući tome što su podaci o prodaji sortirani rastuće po datumu, podeljen na set za trening (koji obuhvata prvih 19 meseci istorijskih podataka) i validacioni set (koji obuhvata poslednjih 5 meseci), čime je osigurana adekvatna vremenska validacija modela na novijim, neviđenim podacima.

Performanse su merene korišćenjem 3 standardne metrike Accuracy, ROC - AUC, F1 score i njihova komparativna analiza data je u nastavku. Kao najrelevantnija metrika za ovaj problem odabrana je F1 score, jer meri kvalitet predikcija po klasama, a ne samo ukupan procenat tačnih pogodaka.

Model	Accuracy	F1-score	ROC-AUC	Preciznost (0)	Recall (0)	F1 (0)	Preciznost (1)	Recall (1)	F1 (1)
Decision Tree	0.827	0.849	0.909	0.73	0.87	0.80	0.91	0.80	0.85
Random Forest	0.855	0.883	0.928	0.83	0.79	0.81	0.87	0.89	0.88
AdaBoost	0.806	0.849	0.884	0.80	0.66	0.73	0.81	0.90	0.85

Decision Tree model daje dobre rezultate i uspešno razlikuje klase, ali iako ima visoku preciznost, povremeno propušta pozitivne slučajeve. Ono sto je dodatno odradjeno je vizuelizacija drveta, gde je prvi I najbitniji atribut bio Promo. Dalje se drvo granalo na osnovu DayOfWeek i toga da li je u pitanju nedelja ili ne (prodavnice zatvorene i nema prodaja). Ovi prediktori su ključni za segmentiranje prodaje na najvišem nivou, potvrđujući da su aktivnost promocije i dan u nedelji izuzetno bitni za prodaju.

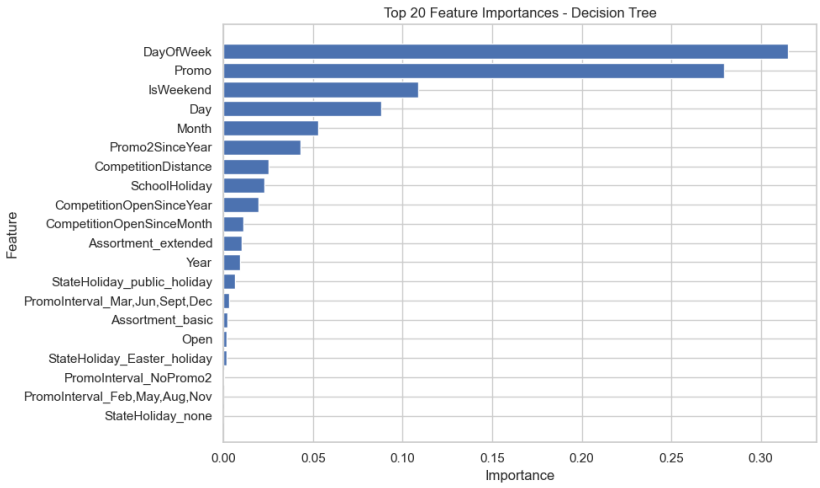


Image 4 - Grafik Feature importances

Random Forest postiže najbolje rezultate među svim modelima, sa stabilnim balansom između preciznosti i odziva. Zbog kombinovanja više stabala, daje robusne i pouzdane predikcije uz najmanje grešaka.

AdaBoost pokazuje nešto nižu tačnost, ali vrlo dobar odziv za pozitivnu klasu. Ovaj model efikasno prepoznaje većinu pozitivnih primera, iako pravi više lažno pozitivnih predikcija u odnosu na Random Forest.

5.2 Vizuelizacija rezultata

ROC kriva (Receiver Operating Characteristic Curve) i njena metrika AUC (Area Under the Curve) predstavljaju standardne alate za procenu kvaliteta modela klasifikacije. Visoka vrednost AUC (blizu 1) ukazuje na odličan diskriminacioni kapacitet modela.

Priloženi grafikoni prikazuju ROC krive za razvijene boosting i bagging modele. Decision tree (AUC = 0.9), Random Forest (AUC = 0.93) i AdaBoost (AUC = 0.88) demonstriraju vrlo snažne performanse, s obzirom da su sve krive daleko iznad dijagonalne linije nasumičnog pogađanja. Ovi rezultati potvrđuju da su odabrani modeli mašinskog učenja izuzetno efikasni za zadatak klasifikacije na osnovu prodaje.

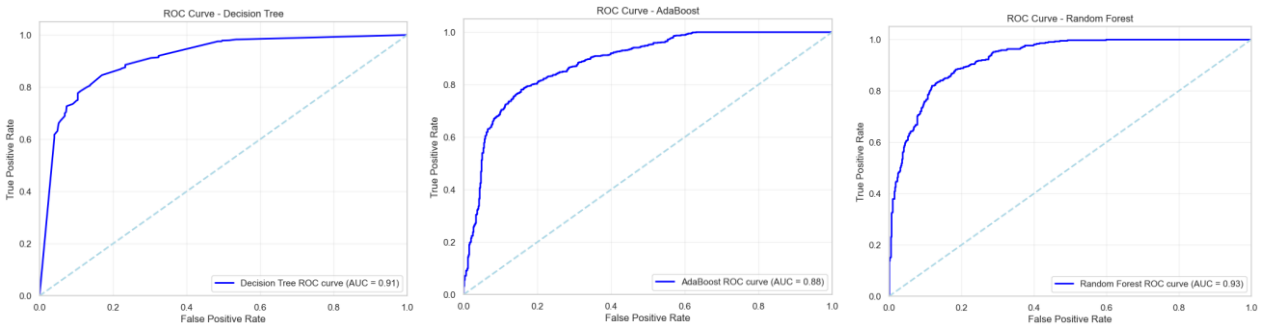


Image 5 - ROC krive

6. Primena modela

Kao finalni korak, vrši se finalno predviđanje prodaje koristeći najbolji obučeni **Random Forest** model na testnom skupu, nakon čega se kreira tabela sa tim predviđenim vrednostima i ID brojevima, spreman za podnošenje rezultata.

7. Zaključak i budući rad

U ovom radu smo transformisali problem prognoze prodaje u zadatak binarne klasifikacije i evaluirali performanse tri *ensemble* modela: Random Forest, AdaBoost i Drvo Odluke. Analiza podataka je potvrdila da su aktivnost promocije (Promo) i dan u nedelji (DayOfWeek) dominantni faktori koji diktiraju da li će prodaja biti klasifikovana kao Visoka ili Niska. Korišćenjem ROC krive, utvrdili smo da je Random Forest model pokazao najbolje rezultate sa AUC vrednostima iznad 0.90.

Za dalji napredak, preporučuje se uvođenje i testiranje modela poput XGBoost i LightGBM, jer su kao napredne *Gradient Boosting* metode, obično jos precizniji. Takodje, bilo bi dobro proširenje skupa ulaznih podataka dodatnim spoljnim (makroekonomskim/demografskim) informacijama i temeljnu proveru netačnih negativnih predviđanja kako bi se smanjio propust poslovnih prilika.

8. Literatura

Rossmann Store Sales. (n.d.). Kaggle. https://www.kaggle.com/competitions/rossmann-store-sales/data

Schapire, R. E. (2013). Explaining AdaBoost. In *Springer eBooks* (pp. 37–52). https://doi.org/10.1007/978-3-642-41136-6_5

Elenapetrova. (2017, July 19). *Time Series Analysis and Forecasts with Prophet*. Kaggle. https://www.kaggle.com/code/elenapetrova/time-series-analysis-and-forecasts-with-prophet