UNIVERZITET U NOVOM SADU
PRIRODNO-MATEMATIČKI FAKULTET
DEPARTMAN ZA MATEMATIKU I
INFORMATIKU

Predikcija otkazivanja Uber vožnji pomoću PySpark-a

-Seminarski rad iz predmeta Analiza Velikih Podataka-

Marko Mirković, 72m/24

Novi Sad, oktobar 2025.

Sadržaj

1 Uvod

U poslednjih nekoliko godina, „**ride-hailing**“ servisi kao što su **Uber**, **Lyft** i slični su transformisali urbani prevoz širom sveta. Ti servisi ne samo da olakšavaju pristup prevozu korisnicima, već i stvaraju **ogromne količine podataka** – o putovanjima, vremenima polaska i dolaska, lokacijama, broju vožnji otkazivanjima i drugim karakteristikama.

Jedan od važnih problema u radu ovih servisa je **otkazivanje vožnji**. Otkazivanje od strane korisnika ili vozača ima više neželjenih efekata: gubitak prihoda, smanjenje korisničkog zadovoljstva, manje efikasno korišćenje vozača i vozila, kao i povećanje troškova operacija. Zato je veoma korisno ako na osnovu istorijskih podataka možemo predvideti da li će neka vožnja biti otkazana.

U ovom radu obrađujemo Uber dataset sa **Kaggle-**a, koji ćemo učitati, analizirati, očistiti, transformisati i konstruisaćemo model za predikciju verovatnoće otkazivanja vožnje. Za obradu svega navedenog koristimo **Apache Spark sa PySpark interfejsom** unutar **Google Colab** okruženja, što nam omogućava skalabilnost i efikasnost. Na kraju ćemo evaluirati naše modele kako bi odredili koji model ima najbolje performanse.

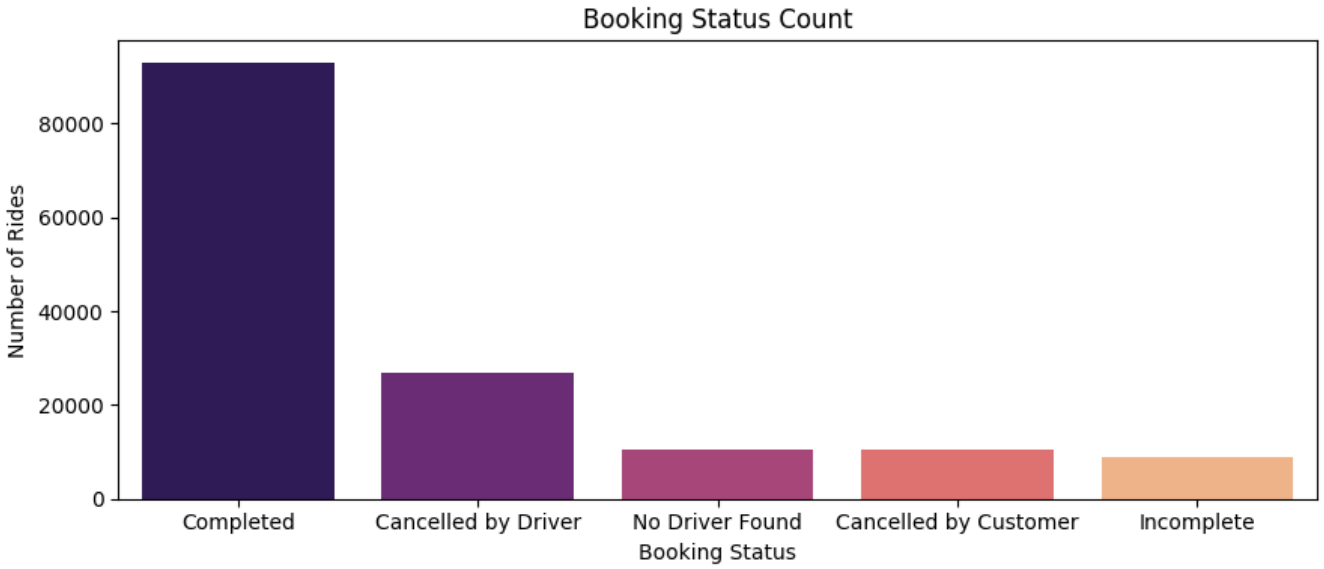
2 Podaci

2.1 Skup podataka

Korišćeni skup podataka nosi naziv **Uber Ride Analytics Dataset 2024** i obuhvata informacije o vožnjama realizovanim putem Uber servisa u Indiji. Dataset sadrži ukupno **148 770 instanci**, pri čemu svaka instanca predstavlja pojedinačnu vožnju različitih tipova vozila. Od ukupnog broja vožnji, **65,96% je uspešno realizovano**, dok je **25% vožnji otkazano** — od čega je **19,15% otkazano od strane korisnika**, a **7,45% od strane vozača**.

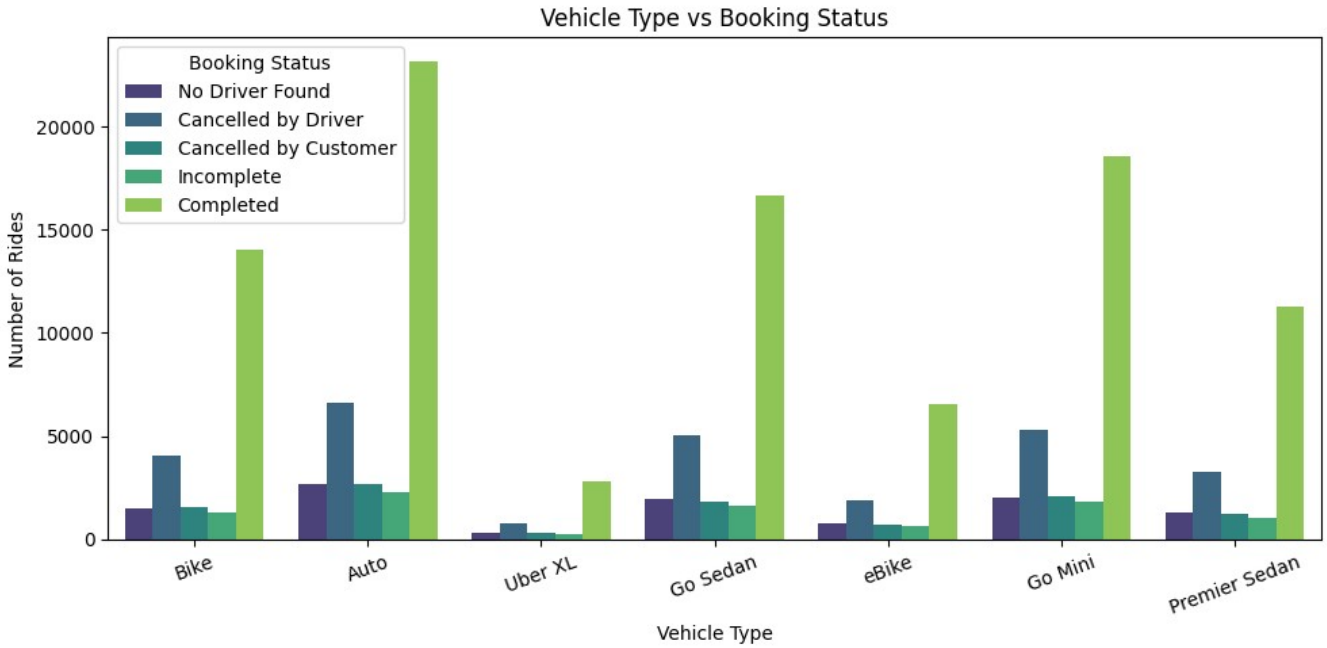
Podaci uključuju različite karakteristike svake vožnje, kao što su **dužina puta**, **ukupna cena**, **razlog otkazivanja** (ukoliko je do njega došlo), kao i **ocene korisnika i vozača**. Takođe, u datasetu su prisutne informacije o **metodu plaćanja** i drugim faktorima koji mogu uticati na tok i ishod vožnje.

2.2 Analiza podataka



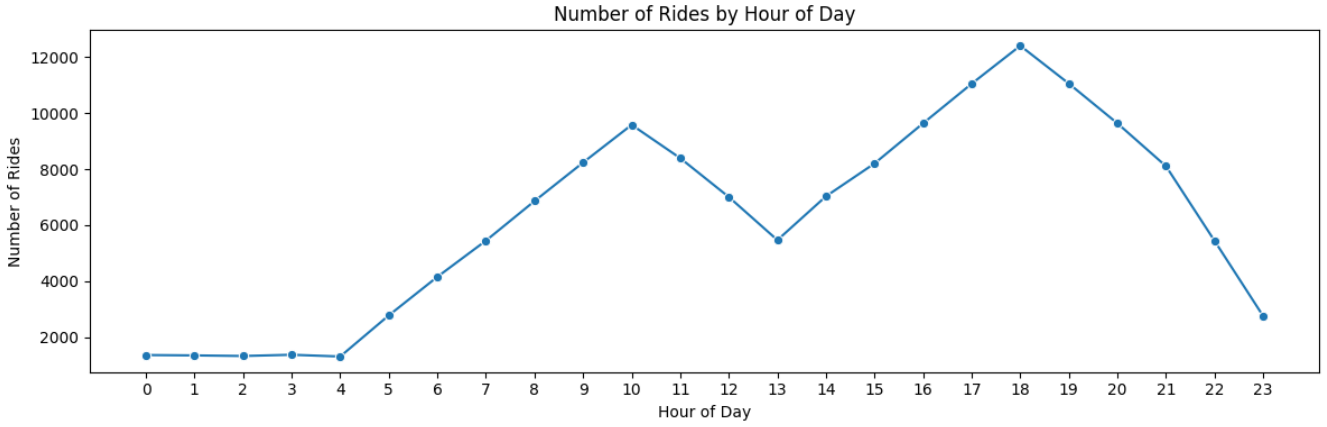
Slika 1: Raspodela statusa vožnji

Vidimo da je većina vožnji u skupu podataka zapravo uspešna, dok je manji broj otkazan ili neuspešan.



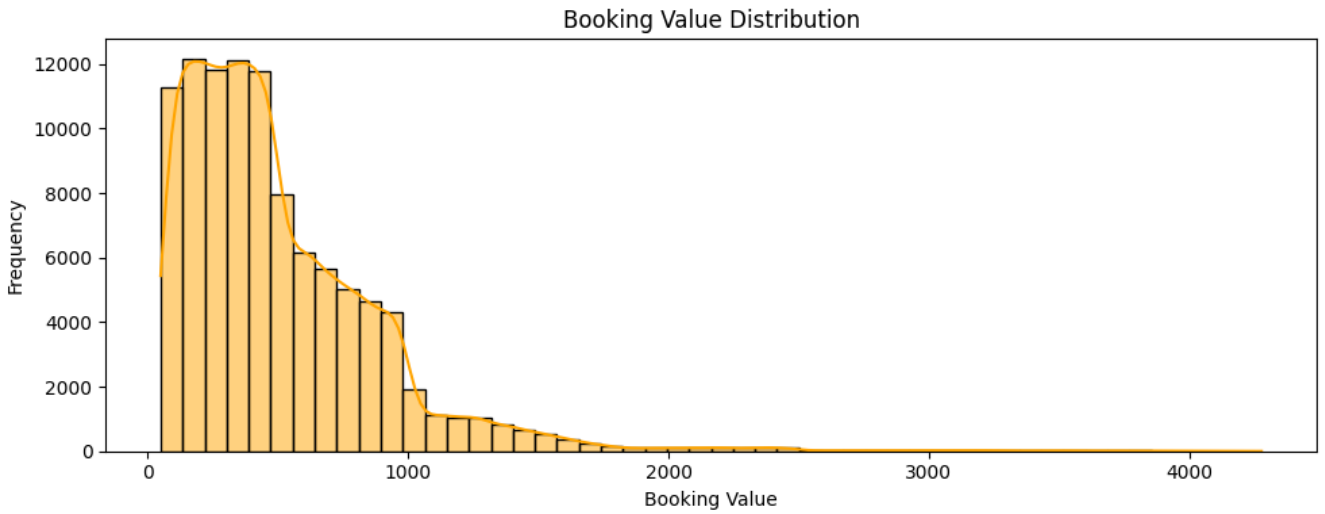
Slika 2: Odnos tipa vozila i statusa vožnje

Vidimo da najveći broj vožnji (samim tim i uspešnih vožnji) se izvodi sa tipom „**Auto**“, dok ostali tipovi blsko prate, osim „**Uber XL**“ koji je noviji i manje rasprostranjen tip vozila, kao i „**eBike**“.



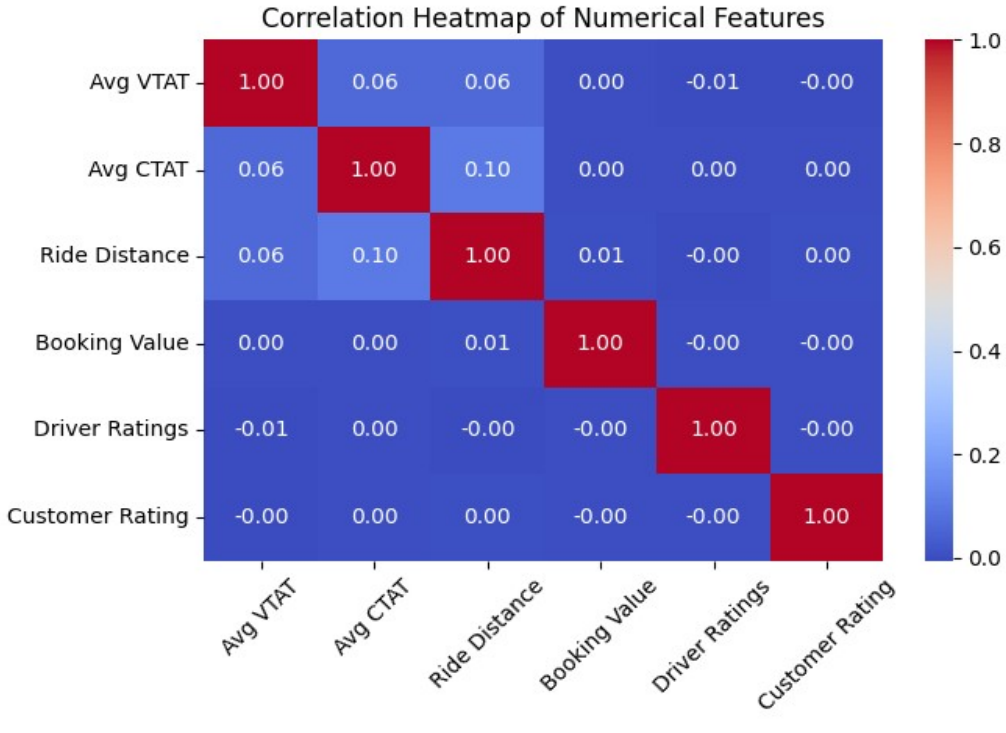
Slika 3: Broj vožnji u odnosu na period dana

Vidimo da broj vožnji ima dva peak-a, jedan u **jutranjim** časovima i drugi u **popodnevnim**, što je skroz očekivano sa obzirom da ljudi idu u školu ili na posao, a popodne se vraćaju.

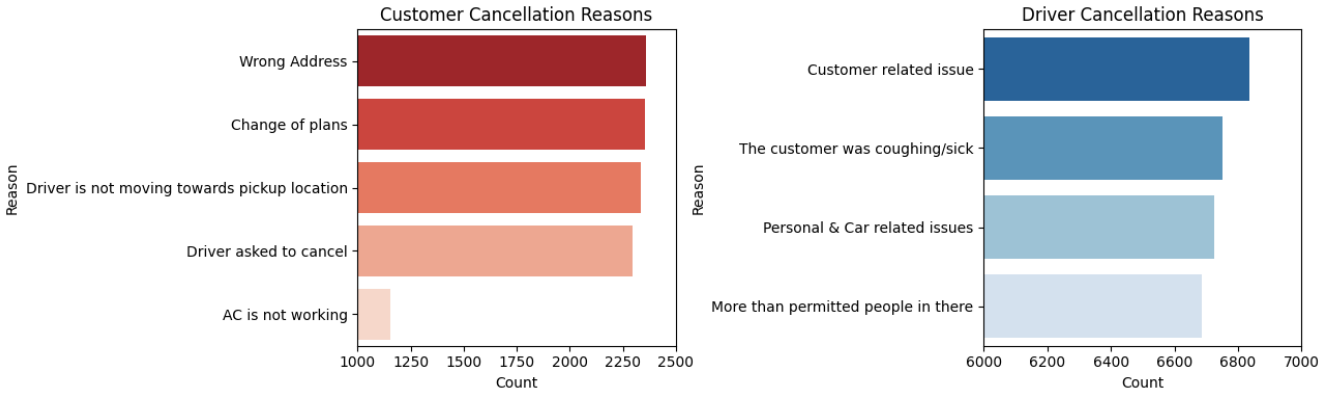


Slika 4: Distribucija cene vožnje

Imamo očekivanu „**power-law**“ raspodelu koja opisuje da je većina vožnji kratko i manje koštaju, dok imamo manji broj vožnji koje su skupe i dugačke.



Vidimo da naši numerički podaci nisu u korelaciji, pa neće biti potrebno dodatno obrađivanje.



Slika 6: Razlozi za otkazivanje vožnji od strane korisnika i vozača

Vidimo najčešće razloge zbog kojih korisnici i vozači otkazuju vožnje, što može da ukaže na potencijalne tačke koje Uber može unaprediti u svojim servisima.

3 Metodologija

Kao što je već napomenuto, za obradu podataka koristimo **PySpark** u kom ćemo pripremiti podatke i trenirati modele mašinskog učenja. Priprema podataka podrazumeva zamenu praznih vrednosti sa prosečnim vrednostima numeričkij tipova podataka kao i indeksiranje i enkodovanje kategoričkih tipova podataka.

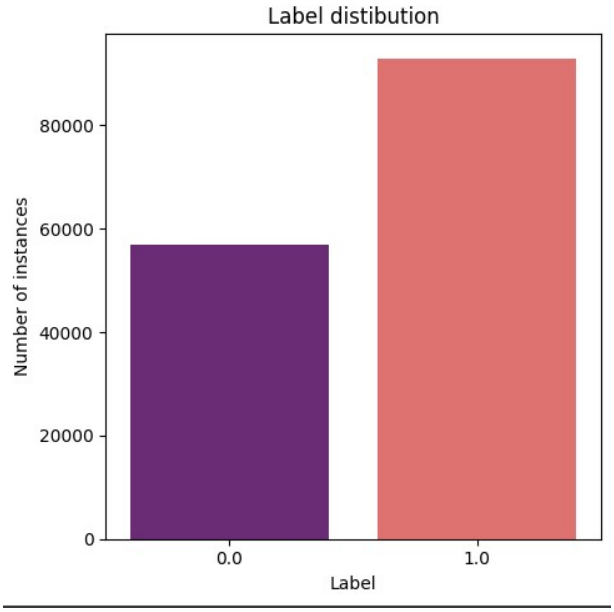
3.1 Modeli mašinskog učenja

Koristimo više različitih modela mašinskog učenja kako bi odabrali koji model najbolje predstavlja naš sistem i ima najbolje performanse nad našim podacima. Modeli koje ćemo koristiti su:

- **Decision Tree** – Model koji se koristi za hijerarhijsku strukturu grananja kako bi donosio odluke, pri čemu se podaci sukcesivno dele prema vrednostima atributa dok se ne postigne čista klasifikacija;
- **Random Forest** – Ansambl metoda koji kombinuje veci broj Decision Tree-a i odluka se donosi glasanjem svih stabala;
- **Naive Bayes** – Probabilistički klasifikator koji se zasniva na Bajesovoj teoremi, koji pretpostavlja verovatnoću da primer pripada određenoj klasi;
- **Gradient Boosted Trees** – Model koji sekvencijalno gradi niz slabijih Decision Tree-a, gde svaki sledeći pokušava da ispravi greške prethodnog;
- **Linear Support Vector Machines** – Linearni model koji pronalazi optimalnu hiperravan koja maksimalno razdvaja klase u prostoru osobina;
- **Neural Networks** – Modeli sastavljeni od slojeva međusobno povezanih neurona (perceptrona) koji kroz proces učenja podešavaju težine veza kako bi prepoznali složene obrasce i odnose u podacima.

Sve ove modele koristimo sa osnovnim parametrima.

3.2 Evaluacija modela



Slika 7: Distribucija uspešnih i neuspešnih vožnji

Na slici 7 možemo videti distribuciju klasa, što znači da je oko dve trećine vožnja uspešno, dok je jedna trećina neuspešna. To nam govori da je skup podataka **nebalansiran** i da trebamo naći **optimalne metrike evaluacije**. Podatke smo podelili na podatke za trening i odvojen skup podataka za evaluaciju kako bi dobili prespristrasan rezultat. Metrike evaluacije koje ćemo koristiti su:

- **Tačnost (Accuracy)** – Predstavlja udeo ispravno klasifikovanih instanci u odnosu na ukupan broj primera, ali može dodati pristrasnosti kod nebalansiranih skupova podataka, pa dodajemo još metrika;
- **Weighted Precision (ponderisana preciznost)** – Predstavlja koliko su predviđanja pozitivne klase tačna, pri čemu se rezultat računa kao ponderisana prosečna vrednosti preciznosti svih klasa, uzimajući u obzir broj instanci po klasama.
- **Weighted Recall (ponderisani odziv)** – Predstavlja koliko dobro model prepoznaje stvarne pozitivne primere u svim klasama, pri čemu se vrednosti ponderišu prema veličini svake klase.
- **F1** – Harmonijska sredina između preciznosti i odziva, koristi se kad je potrebno postići ravnotežu između tačnosti pozitivnih predikcija i sposobnosti modela da prepozna sve relevatne primere.

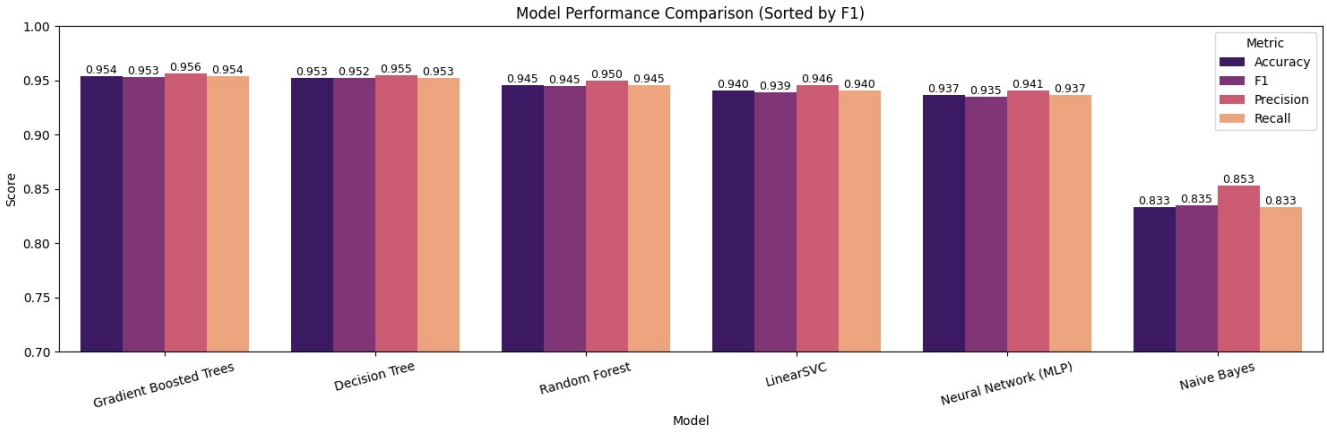
4 Rezultati

Kao što je napomenuto, evaluaciju smo radili na odvojenom skupu podataka, kako ne bi uključili dodatne pristrasnosti modela.

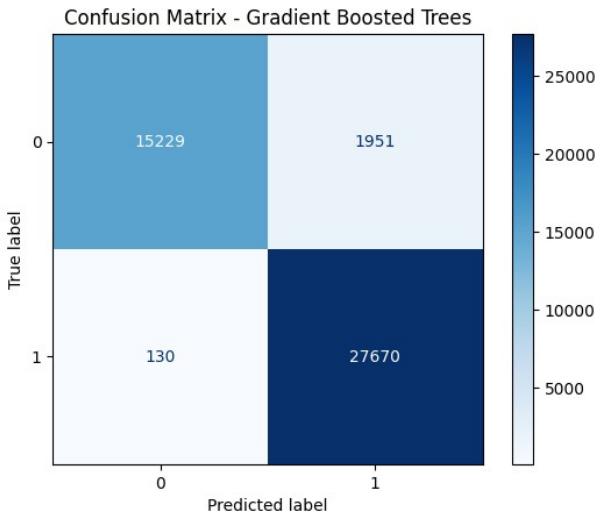
Model	Accuracy	F1	Precision	Recall
Gradient Boosted Trees	0.953735	0.953174	0.956059	0.953735
Decision Tree	0.952735	0.952180	0.954872	0.952735
Random Forest	0.945487	0.944571	0.949654	0.945487
LinearSVC	0.940462	0.939314	0.945694	0.940462
Neural Network (MLP)	0.936505	0.935373	0.940795	0.936505
Naive Bayes	0.832904	0.835196	0.852803	0.832904

Tabela 1: Vidimo vrednosti metrika za svaki model, evaluirano na test skupu, sortirano po F1

Kao što vidimo na tabeli 1, najbolje performanse postiže **Gradient Boosting Tree**. Ali većina modela ima slične performanse, osim „**Naive Bayes-a**“ koji ima malo niže performanse u odnosu na ostale.



Slika 8: Grafikon metrika za modele, sortiran po F1



Slika 9: Matrica konfuzije za najbolji model

Takođe na matrici konfuzije vidimo da je najbolji model po **F1** metrici na test skupu „**Gradient Boosted Trees**“. Vidimo da imamo 1951 **lažno pozitivanih** i 130 **lažno negativnih** primera. U našem slučaju, to je skroz u redu jer smo i dalje ostvarili preko **95% tačnosti**, što nam ukazuje da sa velikom verovatnoćom možemo predvideti da li će vožnja biti otkazana.

5 Zaključak

Na osnovu sprovedene analize i testiranja različitih modela mašinskog učenja, najbolji postignuti rezultati postignuti su sa **Gradient Boosted Trees** algoritmom. Ovaj model ovaj model ima bolju sposobnost učenja **nelinearnih odnosa** i bolje radi sa **mešavinom kategoričkih** i **numeričkih** podataka, pa se pokazao kao **najstabilniji** i **najpouzdaniji** prilikom rada sa složenim podacima koji karakterišu uslove Uber servisa.

Dobijeni model se može praktično primeniti u okviru **sistema za rezervaciju vožnji**, gde bi predikcija verovatnoće otkazivanja bila deo procesa dodele vožnje. Na primer, ako model predvidi da postoji visoka verovatnoća otkazivanja, sistem bi mogao da automatski ponudi alternativnog vozača ili drugi tip vozila.

Na ovaj način, integracijom modela u stvari operativni sistem, Uber bi mogao da **smanji broj otkazanih vožnji**, **poveća zadovoljstvo korisnika** i **vozača**, te **poboljša korišćenja resursa**.

U budućem radu, model se može dodatno unaprediti uključivanjem **većeg broja karakteristika**, poput **vremenskih uslova**, **saobraćaja** ili **dostupnosti vozača**.

6 Reference

1. Yash Devladdha, *Uber Ride Analytics Dashboard Dataset (2024)*. Dostupno na: https://www.kaggle.com/datasets/yashdevladdha/uber-ride-analytics-dashboard/data (pristupljeno: oktobar 2025).
2. *Apache Spark Documentation – PySpark API Reference.* Apache Software Foundation, 2024. Dostupno na: https://spark.apache.org/docs/latest/api/python/index.html (pristupljeno: oktobar 2025).