Travel Ticket Cancellation

By: Rayyan Ghaus Rahmat

Hi, I’m Rayyan Ghaus Rahmat

I’m junior data scientist with experience in projects involving Data Analysis, Data Manipulation, Data Visualization, and Machine Learning. I possess the ability to efficiently process data, analyze patterns, and build machine learning models to derive valuable insights.

Projects:

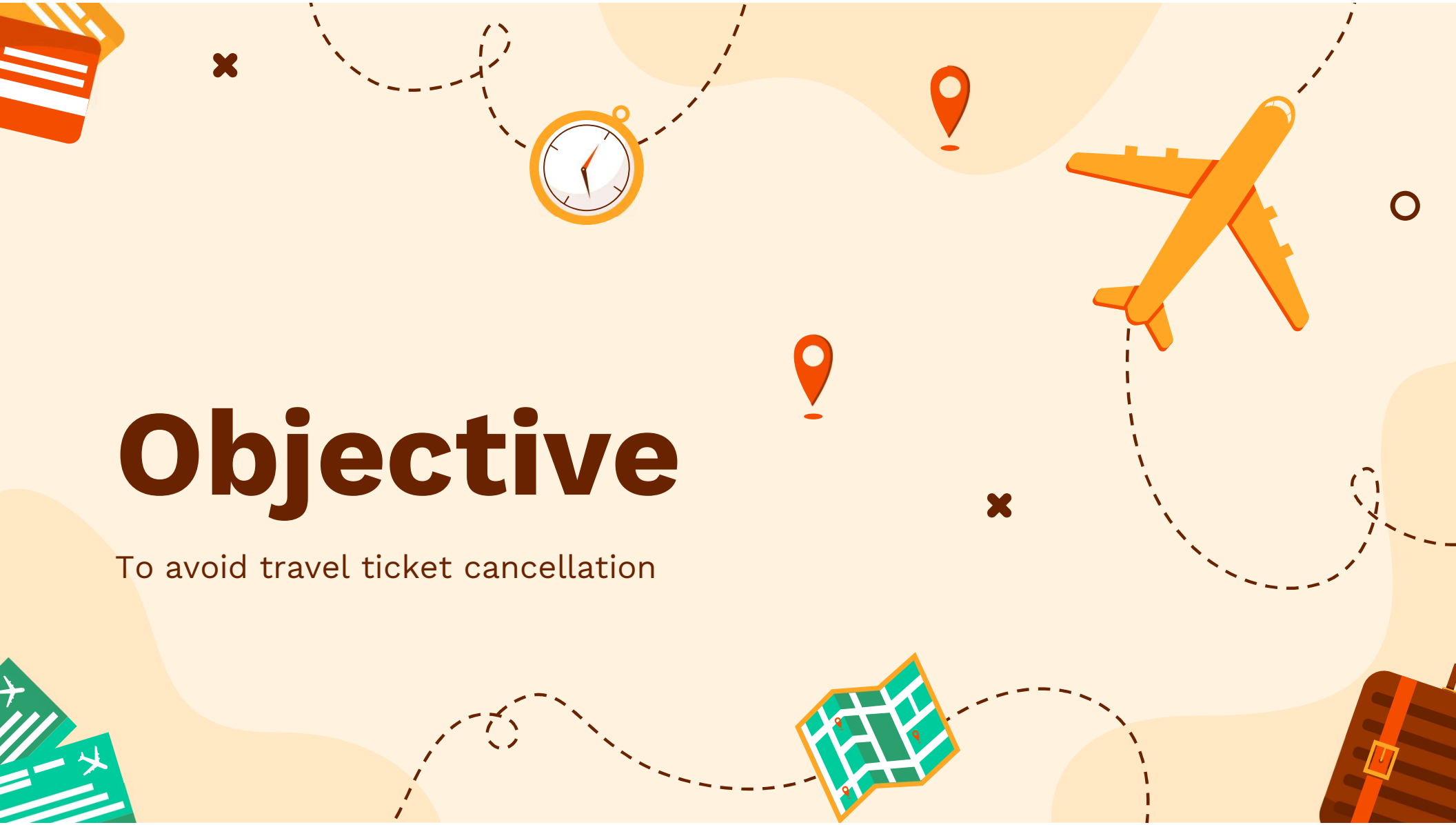
Travel Ticket Cancellation
Credit Risk Classification

More Details:

Linkedin: https://www.linkedin.com/in/rayyan-ghaus-rahmat/
Github: https://github.com/rayyangrahmat
Kaggle: https://www.kaggle.com/rayyaghausrahmat

Objective

To avoid travel ticket cancellation



Outline

01
Business & Data Understanding

02
Data Preprocessing

03
Data Analysis

04
Modelling

05
Conclusion

06
Business Recommendation

Business Understanding

Predicting travel ticket cancellations is crucial for various reasons:

- **Company Profitability**: Anticipating cancellations can significantly impact company profit by optimizing income and enhancing the reputation of the travel agency. This involves strategies to increase income, expand services, and build a larger and better travel brand.
- **Operational Efficiency**: By predicting cancellations, the travel agency can prioritize routes and transportation options that exhibit lower tendencies of ticket cancellations. This leads to improved operational efficiency as resources can be allocated more effectively.
- **Service Regulation**: Understanding the timing of cancellations enables the implementation of effective service regulations. This knowledge can be leveraged to prevent ticket cancellations by creating and enforcing regulations that address common cancellation patterns.

About Dataset

The dataset contains various information about passengers who have registered for a trip through a travel booking website.
This dataset provides valuable insights into passenger travel patterns, booking behavior, and trip cancellations, which can be used for various analyses and predictions in the travel industry.

21 → **13**
Features → **Features**

101017 → **101015**
Rows → **Rows**

For more info: **Kaggle**

Data Understanding

Missing Values

No	Feature	Missing Values	Percentage(%)
1	HashPassportNumber_p	100155	99.15
2	CancelTime	85691	84.83
3	UserID	58474	57.89
4	HashEmail	57933	57.35
5	VehicleClass	38450	38.06

2

Duplicated Rows

2

Numerical Columns

11

Categorical Columns

For more info:
Kaggle

Data Understanding

Drop Effectless Columns

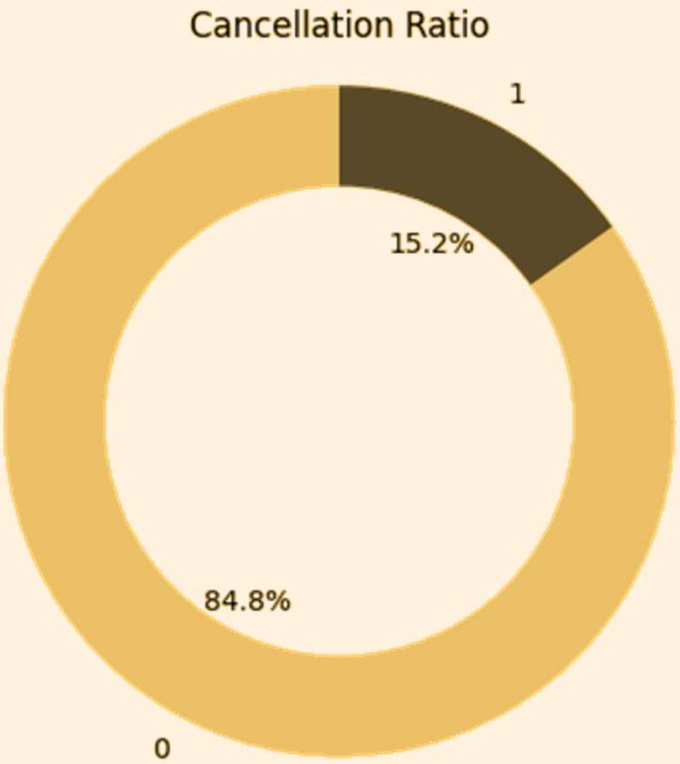
Drop NationalCode, BuyerMobile Because the data is provided in hashed version, it is difficult to process for machine learning and doesn't effects cancellation.
Drop VehicleType because the data is not provided well. The data for VehicleType should be categorical so that insights can be extracted.

BuyerMobile	NationalCode	VehicleType
764974891906	477368495	NaN
27479149496	15987669	NaN
323657282999	667640412	VIP 2+1
169459057632	392476186	ستاره اتوبوسي 3
408595008421	79497837	اسکانیا تک صندلی ۳۱نفره

For more info:
Kaggle

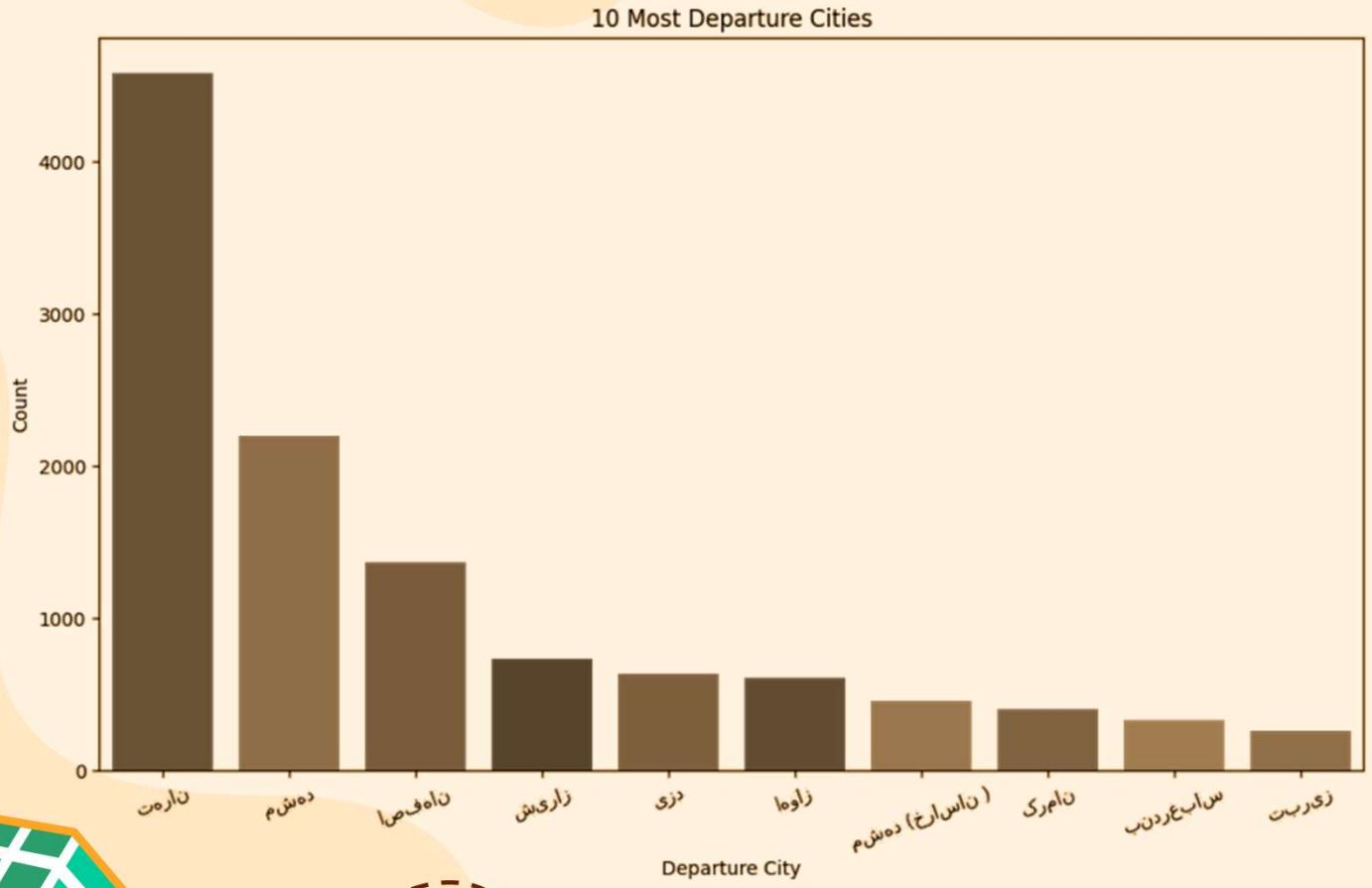

Data Analysis

Cancellation



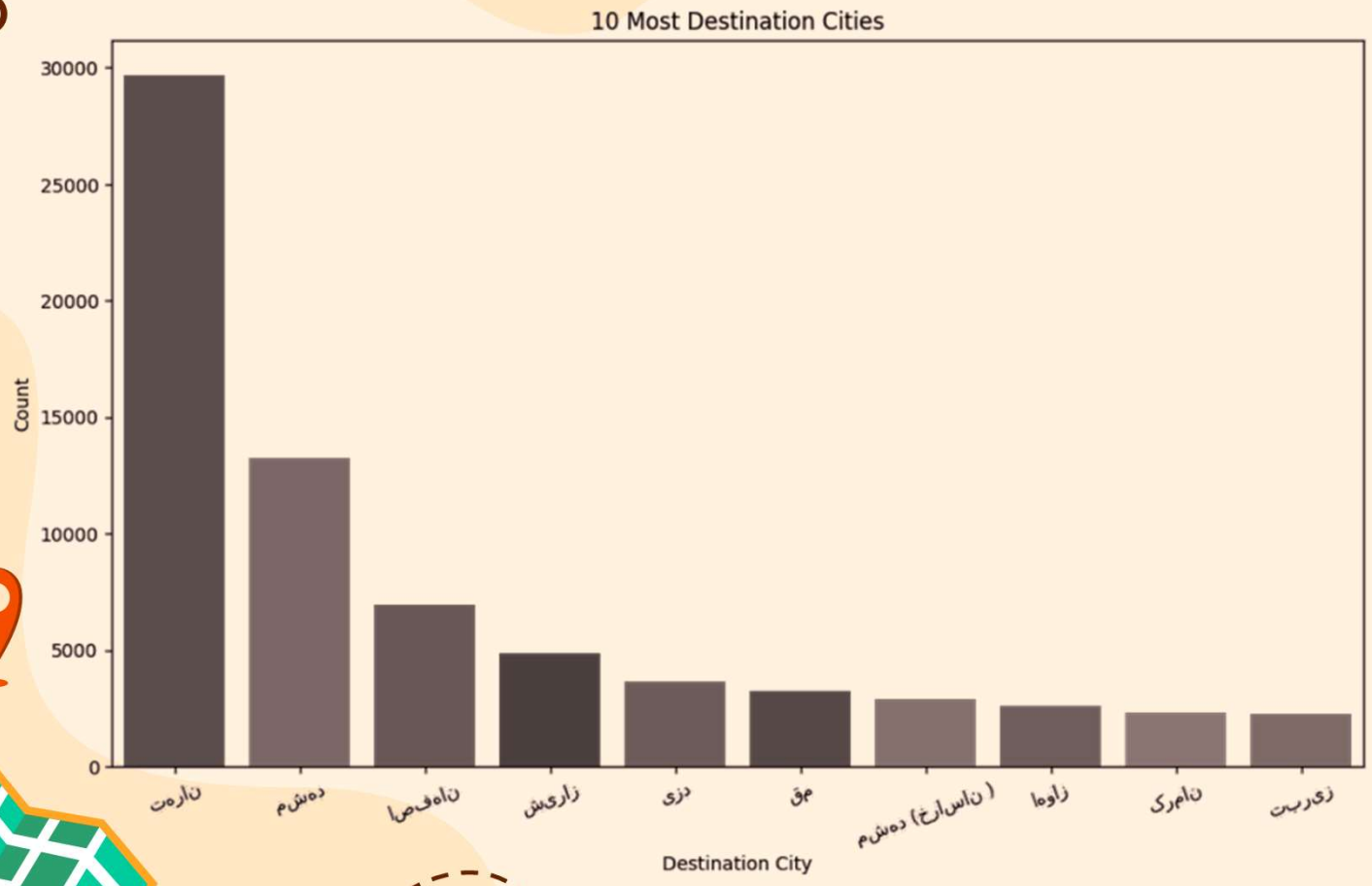
The cancellations ratio is quite skewed or imbalanced. Out of 101,015 rows, only 15.2% resulted in cancellations. Undersampling needs to be performed first to generate balanced data for using it in the machine learning process.

10 Most Departure Cities



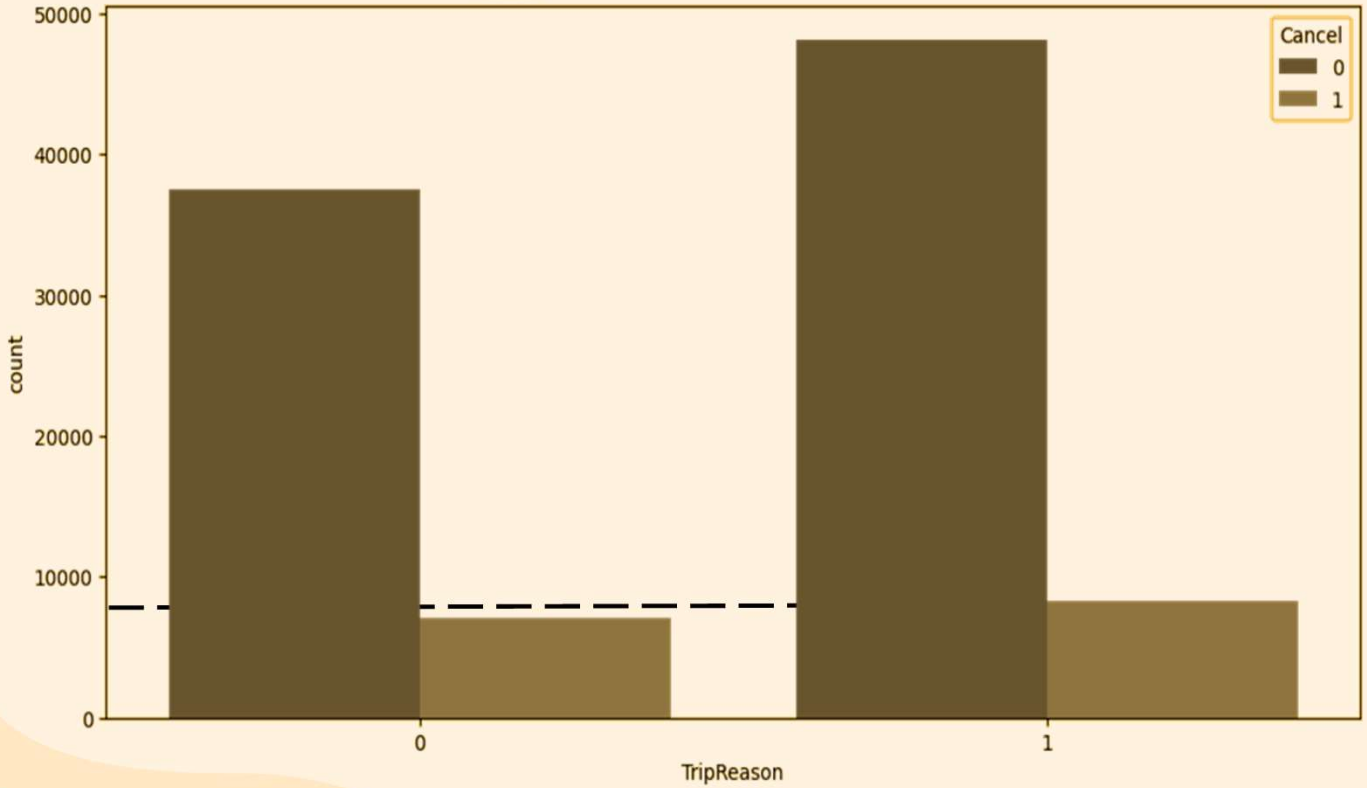
No	City
1	Tehran
2	Mashhad
3	Isfahan
4	Shiraz
5	Yazd
6	Mashhad(Greater Khorasan)
7	Ahvaz
8	Kerman
9	Qom
10	Bandar Abbas

10 Most Destination Cities



No	City
1	Tehran
2	Mashhad
3	Isfahan
4	Shiraz
5	Yazd
6	Qom
7	Mashhad(Greater Khorasan)
8	Ahvaz
9	Kerman
10	Tabriz

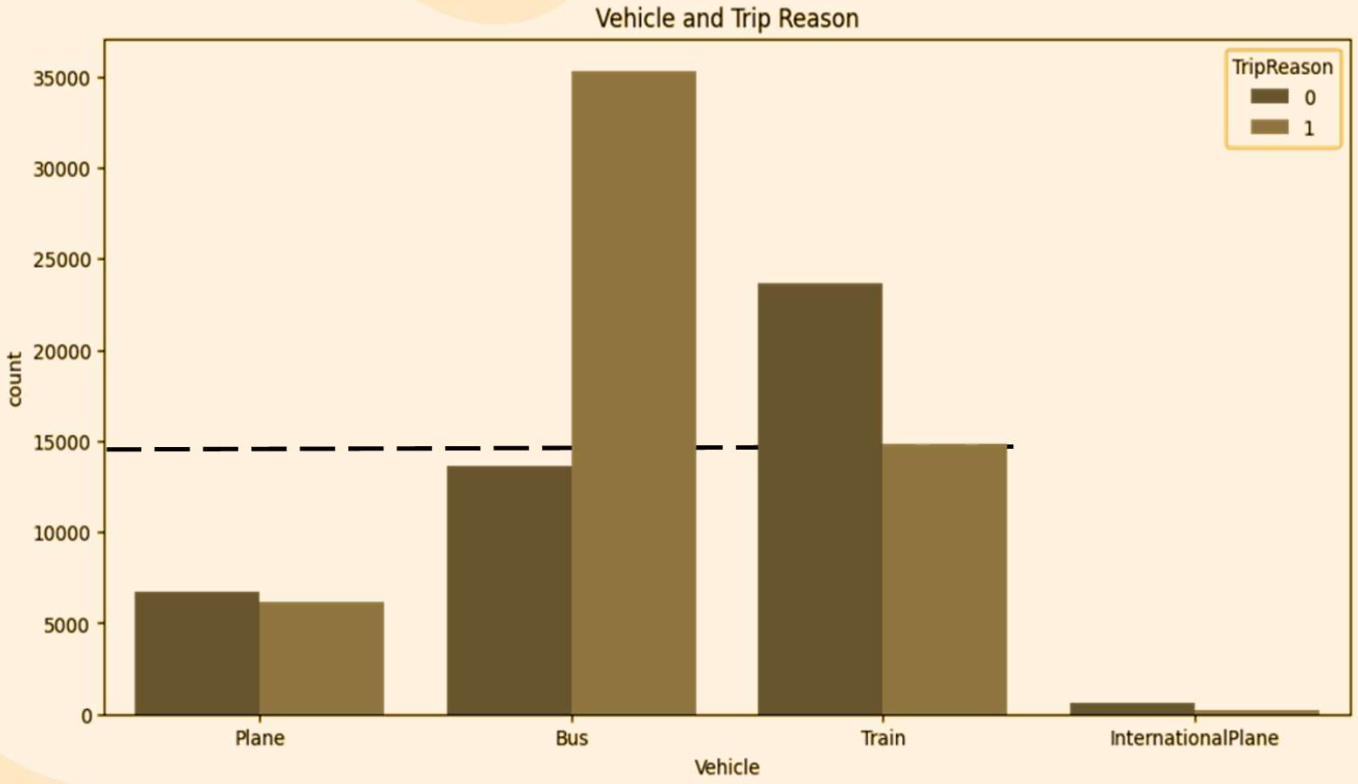
Cancellation and Trip Reason



The reason for working on our customer's trips is more than for international trips. Similarly, after dividing them into cancel and non-cancel.

1: Work
0: International

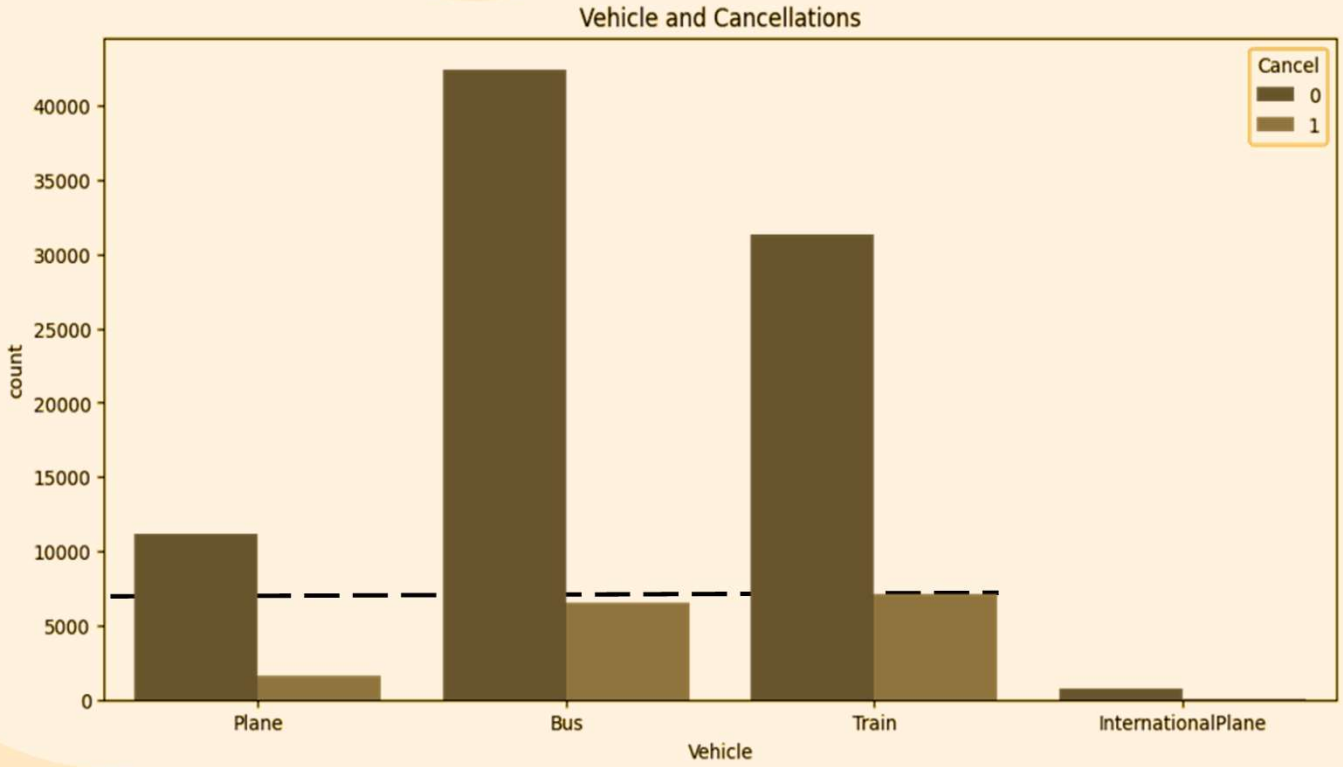
Vehicle and Trip Reason



The most commonly used vehicles are buses and train. For work-related purposes, bus are the most frequently used, while for international purposes, train are the most common.

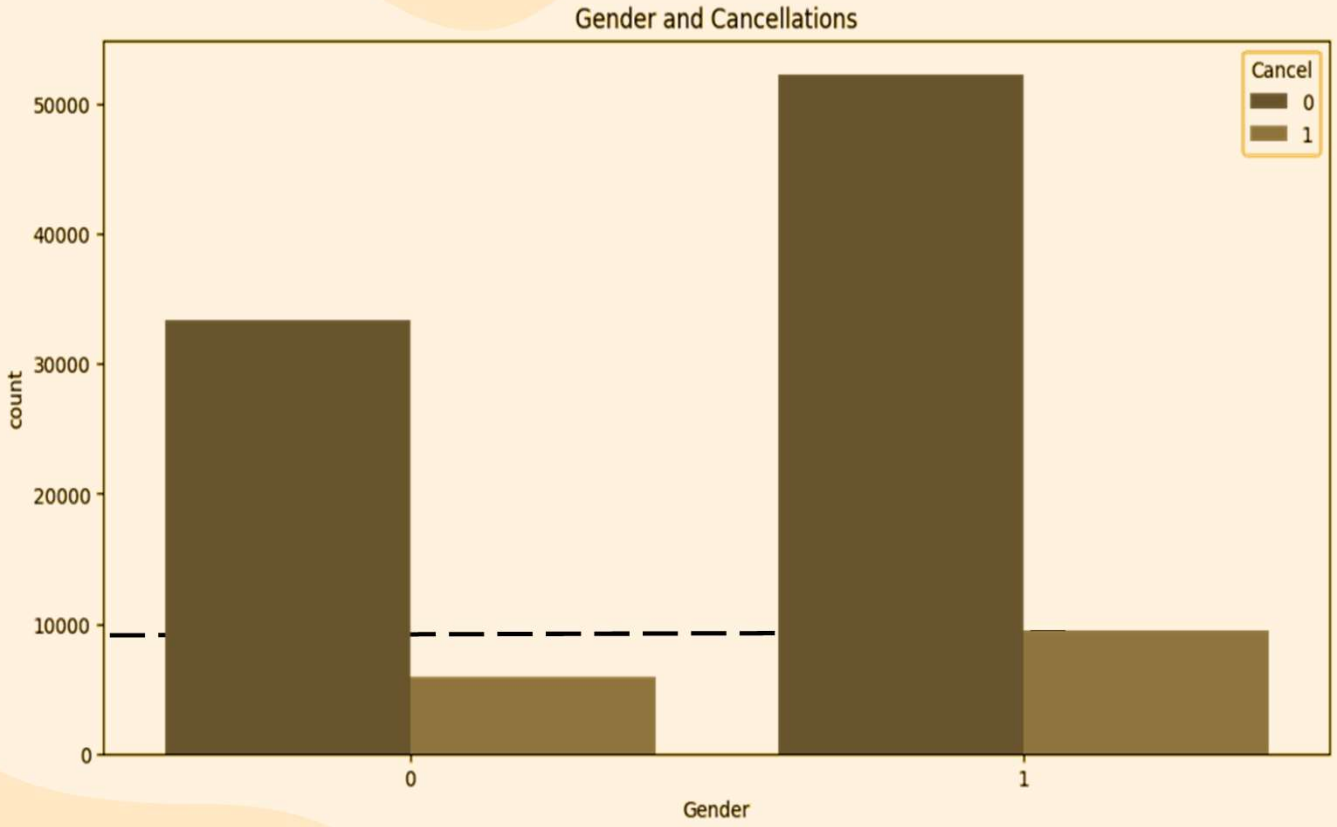
1: Work
0: International

Vehicle and Cancellation



The most frequently used vehicles are bus and train. The highest occurrence of cancellations is with train.

Gender and Cancellation



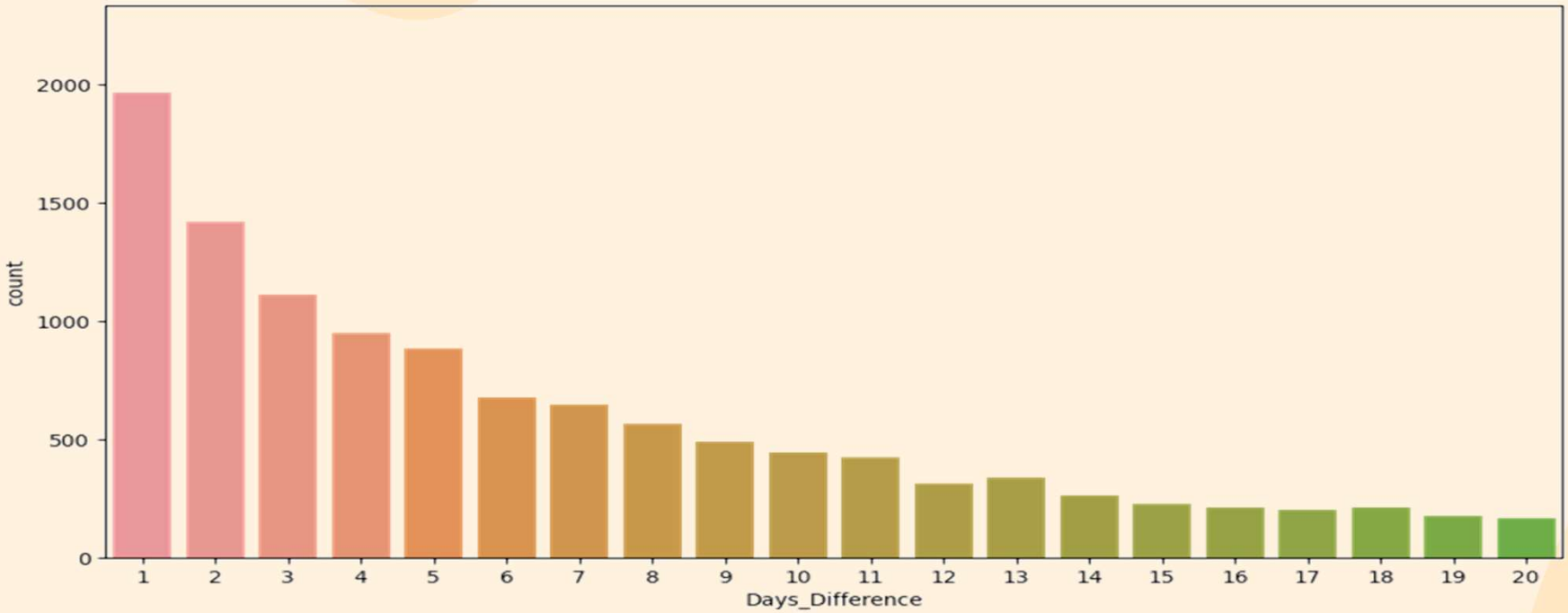
The majority of this company's customers are male. The highest rate of cancellations is among male.

1: Male
0: Female

Monthly Income

There is a slight difference in income, between net income and gross income

Cancellations From Days Difference



The ticket cancellation occurred within a short time frame between the booking time and departure time.

Modelling

Modelling

Split Data

1st Split: Pretrain 80% — Test 20%

2nd Split: Train 80% — Validation 20%

Metric

Baseline:

We use recall as a metric, because the data is imbalanced and we want to avoid false prediction of individuals who are predicted not to cancel but, in reality, they will cancel.

After preprocessing data:

We use accuracy as a metric, because we have handle the imbalanced data.

Modelling (Baseline)

Training Data Results

No.	Model	Accuracy	Precision	Recall	F1 Score
1	XGBoost	0.981439	0.988804	0.889247	0.936387
2	KNN	0.889377	0.758749	0.410391	0.532671
3	Decision Tree	0.973953	0.914722	0.915828	0.915275
4	Random Forest	0.980944	0.968145	0.905759	0.935913

Validation Data Results

No.	Model	Accuracy	Precision	Recall	F1 Score
1	XGBoost	0.981933	0.985961	0.893638	0.937532
2	KNN	0.892244	0.783163	0.400653	0.530110
3	Decision Tree	0.974855	0.919318	0.914519	0.916912
4	Random Forest	0.982379	0.973436	0.908646	0.939926

Modelling (After Preprocessing Data)

Training Data Results

No.	Model	Accuracy	Precision	Recall	F1 Score
1	XGBoost	0.958700	0.962366	0.955030	0.958684
2	KNN	0.629637	0.623873	0.659299	0.641097
3	Decision Tree	0.956979	0.957999	0.956555	0.957086
4	Random Forest	0.961377	0.966128	0.905759	0.961318

Validation Data Results

No.	Model	Accuracy	Precision	Recall	F1 Score
1	XGBoost	0.959009	0.962940	0.954027	0.958462
2	KNN	0.624809	0.616706	0.642394	0.629288
3	Decision Tree	0.953503	0.952542	0.953718	0.953130
4	Random Forest	0.963292	0.969346	0.956186	0.962721

Hyperparameter Tuning

Parameters	Mean Test Score
Max Depth: 2 n_Estimators: 50	1.000000
Max Depth: 1 n_Estimators: 20	0.999427
Max Depth: 2 n_Estimators: 20	0.998282
Max Depth: 2 n_Estimators: 40	0.998187

We are using random forest as the model for hyperparameter tuning because based on the model performance, random forest is the best-performing model. According to the table above, the hyperparameter results in the first row indicate a value of 1.0000, which means it successfully predicted correctly with 100% accuracy.

Conclusion

The dataset consists of 101,017 rows and 21 columns. It is imbalanced and contains missing values, duplicated rows, and outliers. I performed label encoding, one-hot encoding, and changed the data type for the features. After preprocessing, the data has 25 columns and 32,686 rows. I split the dataset into 80% pretraining data and 20% test data, further dividing the pretraining data into training and validation sets.

I chose XGBoost, KNN, Decision Tree, and Random Forest as models to predict travel ticket cancellations. From the baseline model, I obtained results for training and validation that were not significantly different, indicating that the model is suitable for the data, except for the KNN model. The best result from the baseline model was the **Decision Tree**. After preprocessing the data, the results for training and validation were also not significantly different, indicating that the model is suitable for the data. The best result after preprocessing the data was **Random Forest** with higher accuracy than Decision Tree baseline model.

From this research, we can conclude that the performance of tree-based models is better than non-tree-based models. Tree-based models perform well even without preprocessing the data, suggesting that tree-based models are more robust with the data.

Business Recommendation

1. Tightening regulations on booking and cancellation processes aims to reduce cancellations.
2. Providing the option to reschedule for customers intending to cancel their travel tickets is essential.
3. We need to establish fixed departure schedules for each route to enhance company resource efficiency.
4. Offering travel options that allows customers to choose more cost-effective alternatives.
5. Promoting destinations with the highest demand helps decrease the likelihood of ticket cancellations.

Thank You

Contact.
rghausr@gmail.com

kaggle GitHub