Deloitte.

AI Academy Capstone Team 2

March 2024

Meet the Team!

Sabrina Callejo
Solution Analyst
Deloitte Consulting
scallejo@deloitte.com

Syed Ali Irtaza
Solution Analyst
Deloitte Consulting
sirtaza@deloitte.com

Ari Kohl
Analyst
Deloitte Risk & Financial Advisory
arikohl@deloitte.com

Levin Sam
Analyst
Deloitte Risk & Financial Advisory
lesam@deloitte.com

Overview

Introduction

Predicting returns in e-commerce using the Kaggle dataset

[Retailers' average return rate jumps... (cnbc.com)](https://www.cnbc.com)

16.6% of total merchandise returned in 2021, a jump from an average return rate of 10.6% in 2020

Introduction/Business Understanding

What are the factors that influence buying behavior of e-commerce customers to minimize returns of products for businesses?

Our goal: To help the e-commerce market industry gain insights on buying behavior to determine an effective way of increasing revenue.

Focused primarily on determining factors that lead to returns from customers

Dataset

https://www.kaggle.com/datasets/willianoliveiragibin/websites-e-comerce

	accessed_Ffom	age	gender	country	membership	language	returned	pay_method
0	Chrome	28	Female	CA	Normal	English	No	Credit Card
1	Mozilla Firefox	21	Male	AR	Normal	English	No	Debit Card
2	Mozilla Firefox	20	Male	PL	Normal	English	No	Cash
3	Mozilla Firefox	66	Female	IN	Normal	Spanish	No	Credit Card
4	Mozilla Firefox	53	Female	KR	Normal	Spanish	No	Cash

Pre-Processing Steps

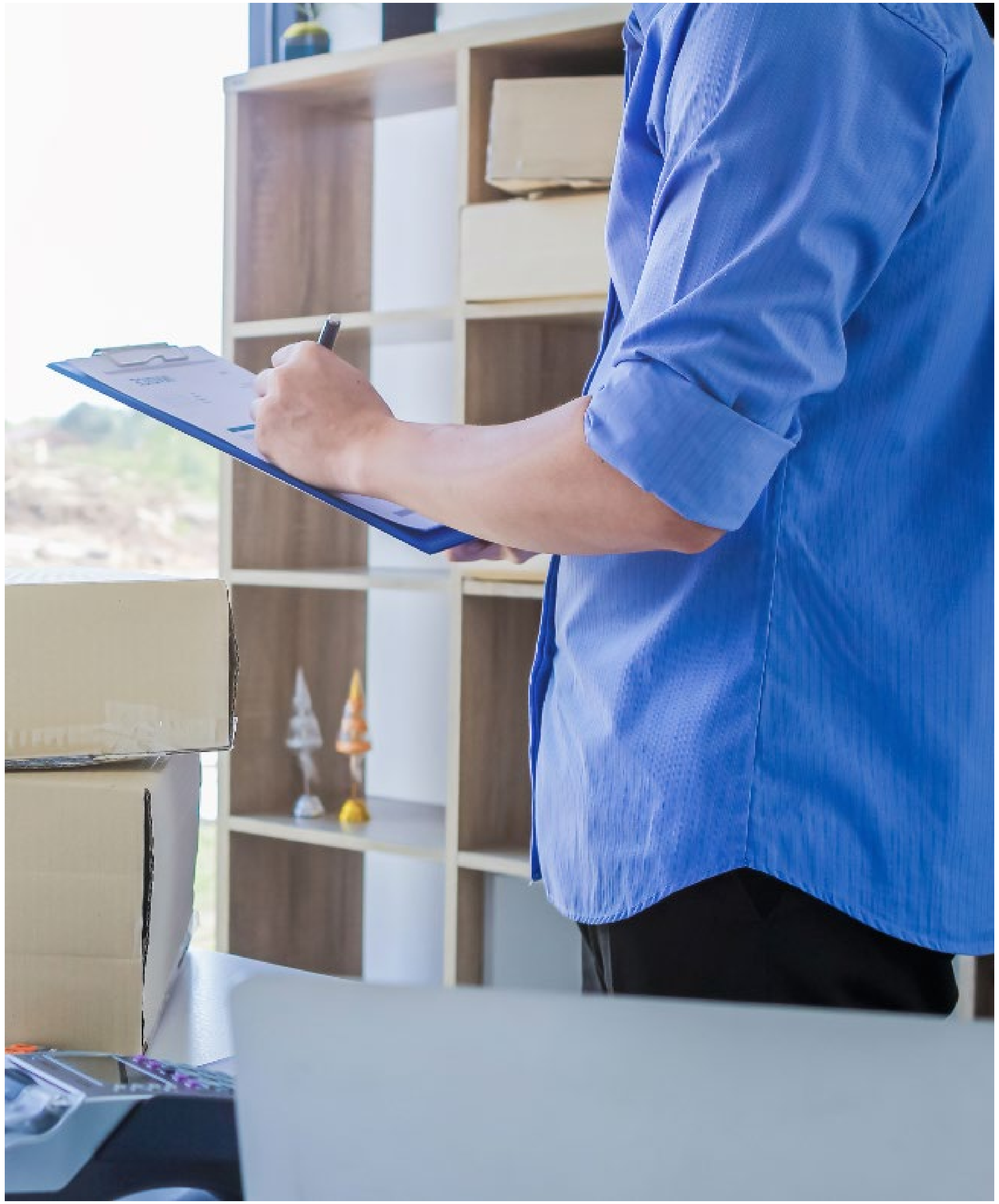
Step 01 – **Drop Irrelevant Columns**

Step 02 – **Null Value Check**

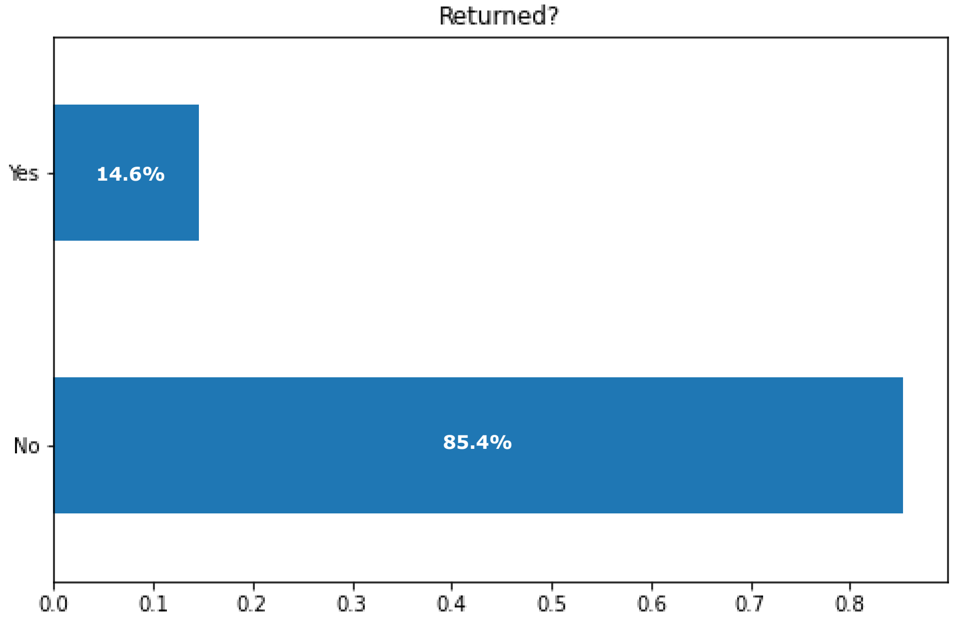
Step 03 – **Handle Placeholder Values**

Step 04 – **Fixed Formatting in Columns**

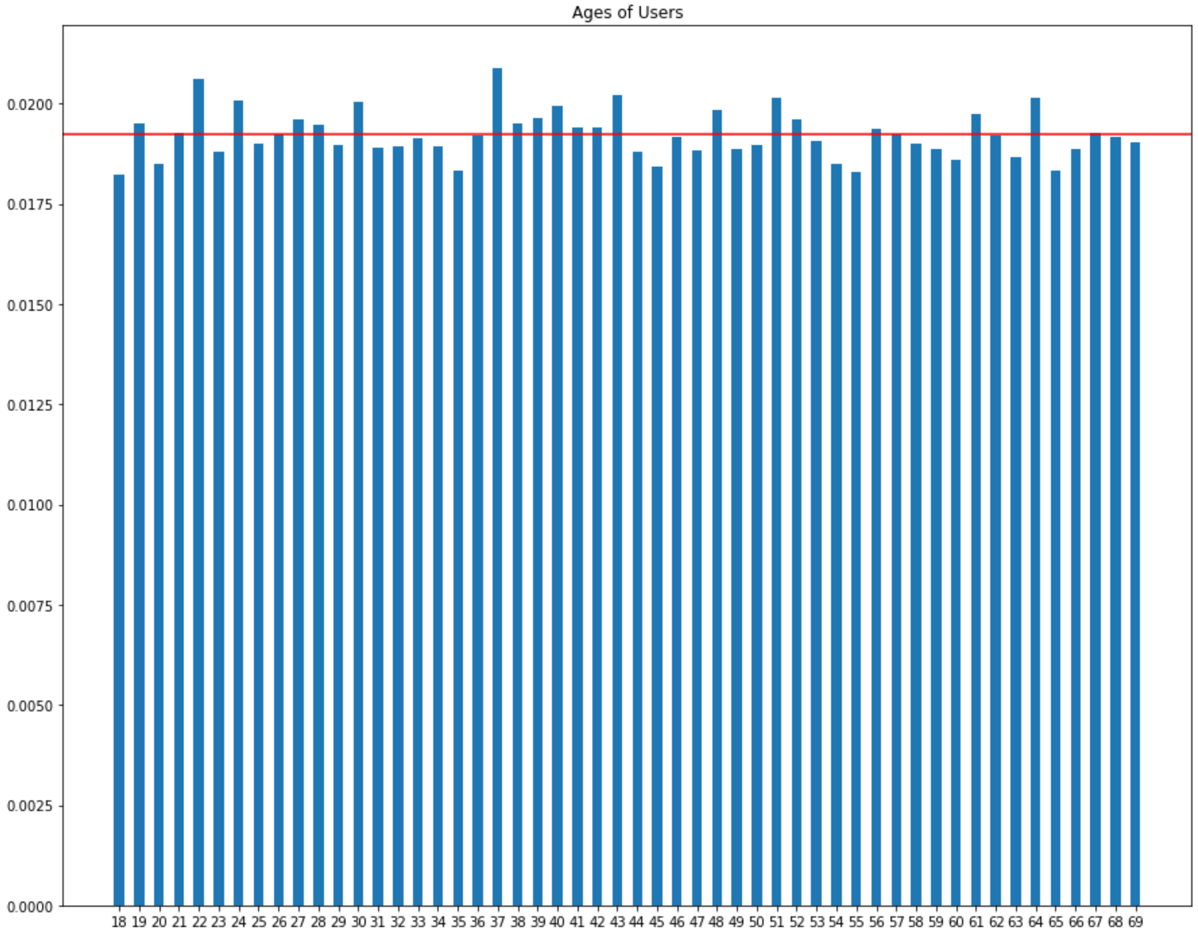
Data Exploration

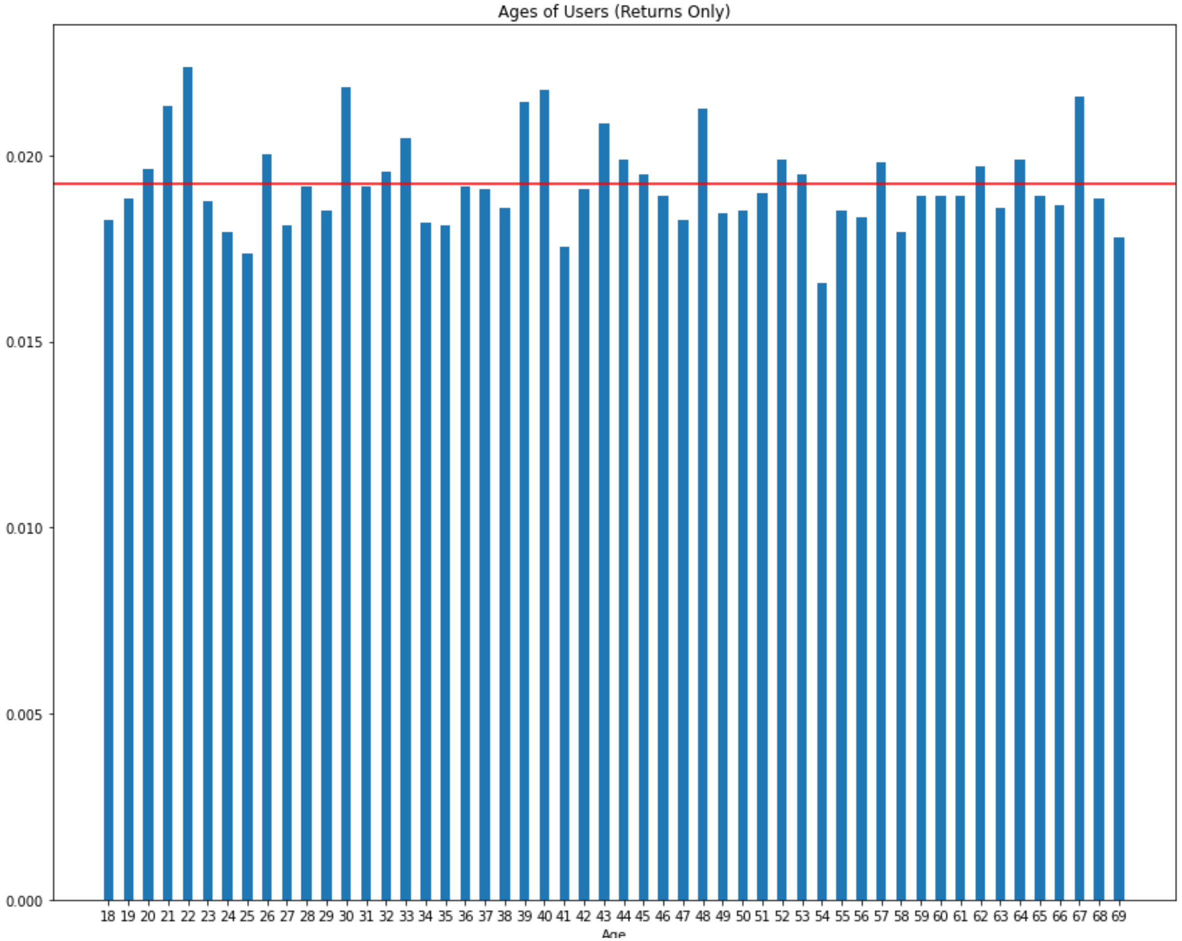


Data Exploration – Target Variable



Data Exploration – Age of Users





Top 5 Ages

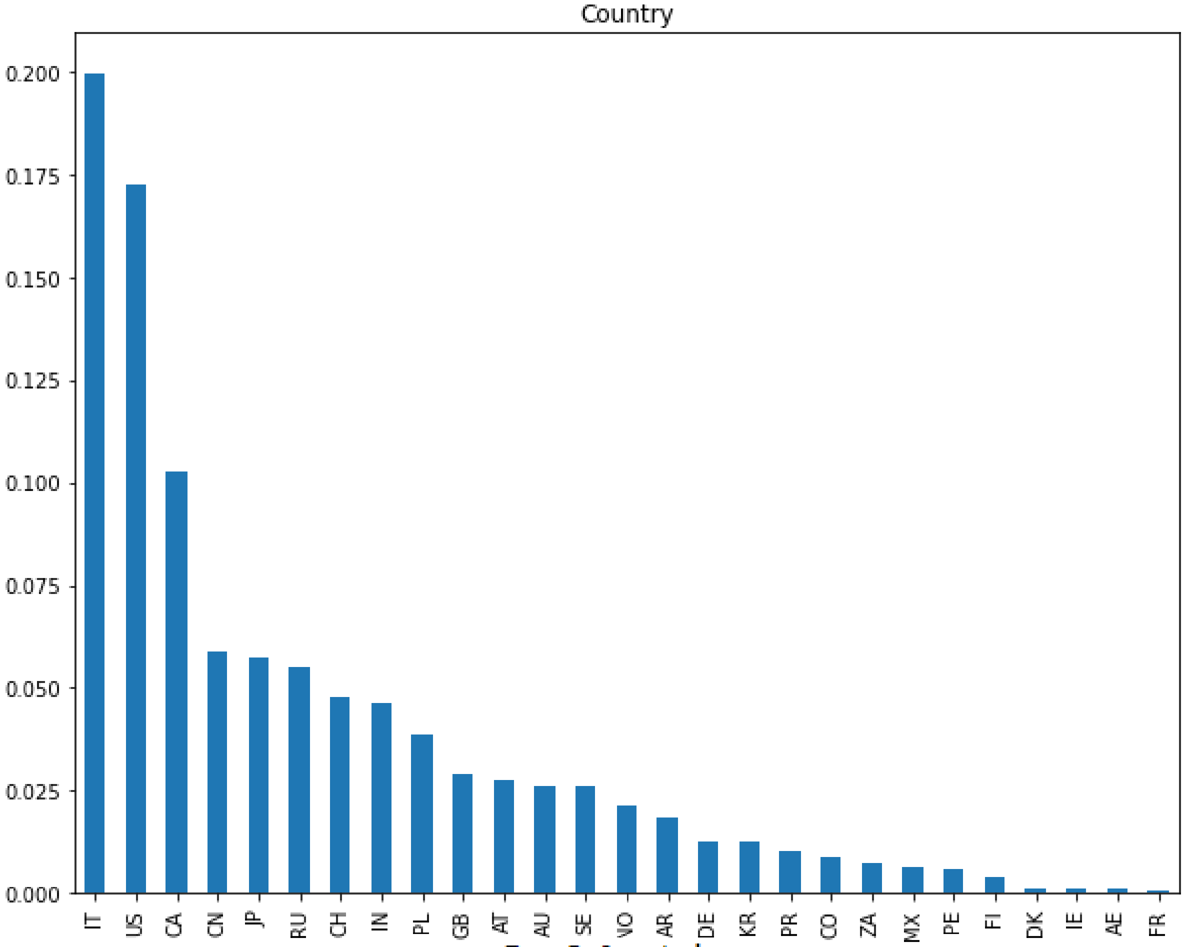
37	2.09%
22	2.06%
43	2.02%
51	2.02%
64	2.02%

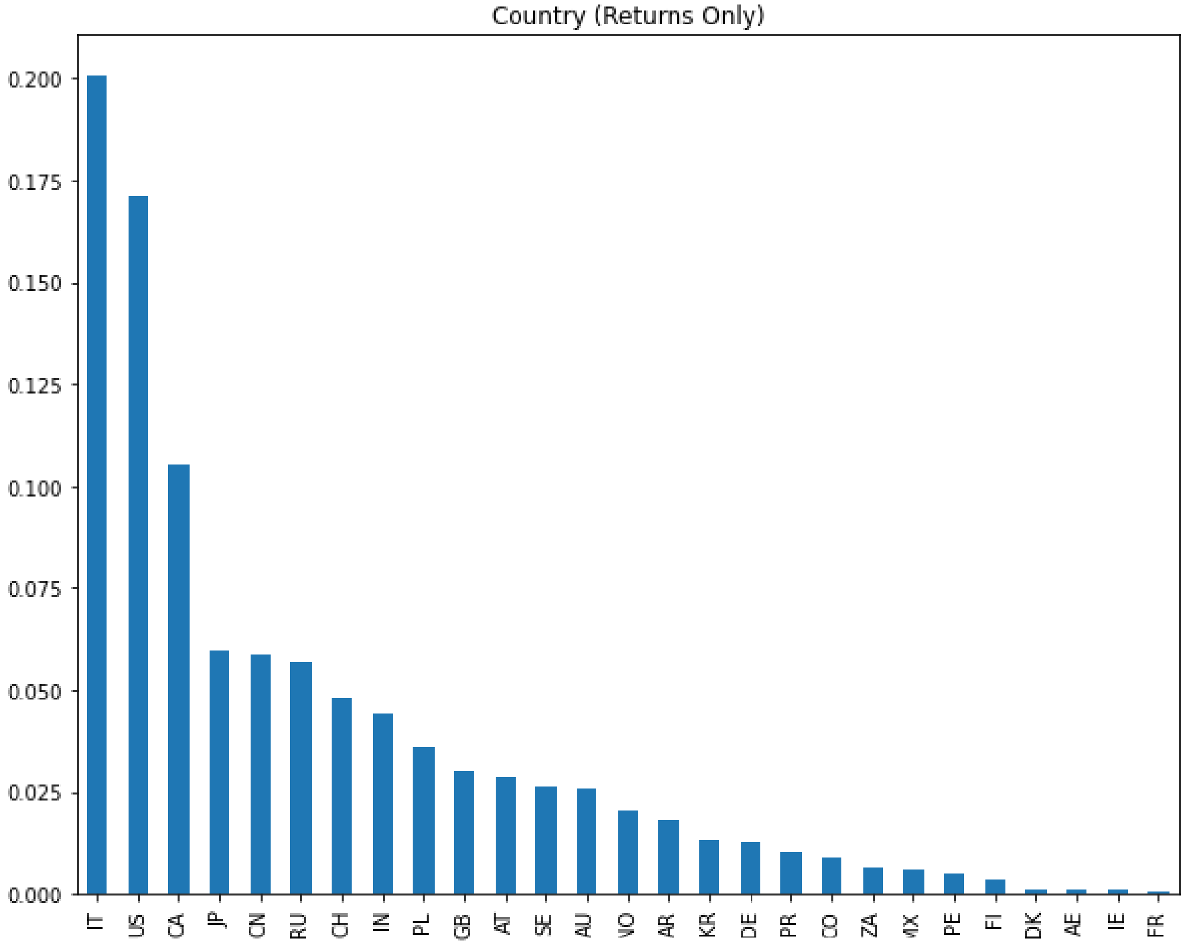
- Similar distributions in the 2 charts
  - Different Ages in the top 5s

Top 5 Ages (Returns Only)

22	2.22%
30	2.18%
40	2.17%
67	2.15%
39	2.14%

Data Exploration - Country





Top 5 Countries

IT 19.9%

US 17.2%

CA 10.3%

CN 5.9%

JP 5.7%

Top 5 Countries (Returns Only)

IT 20.1%

US 17.1%

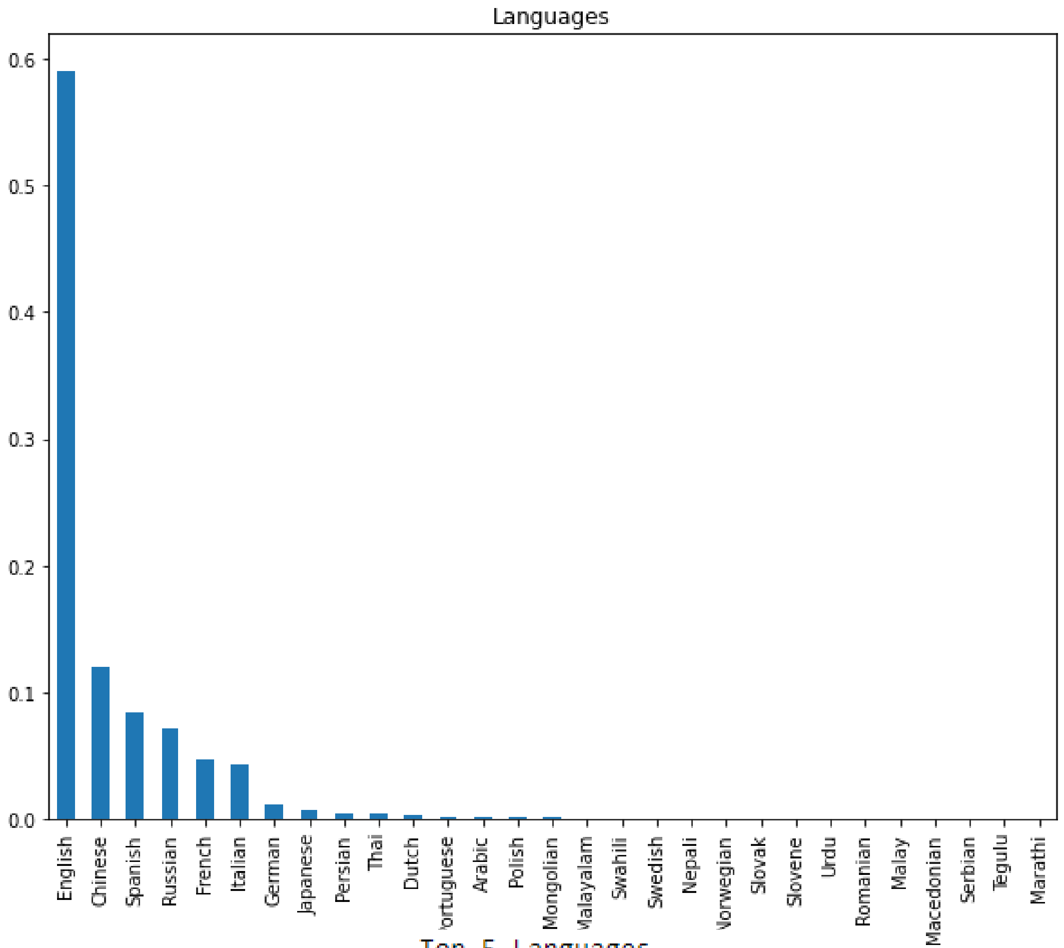
CA 10.5%

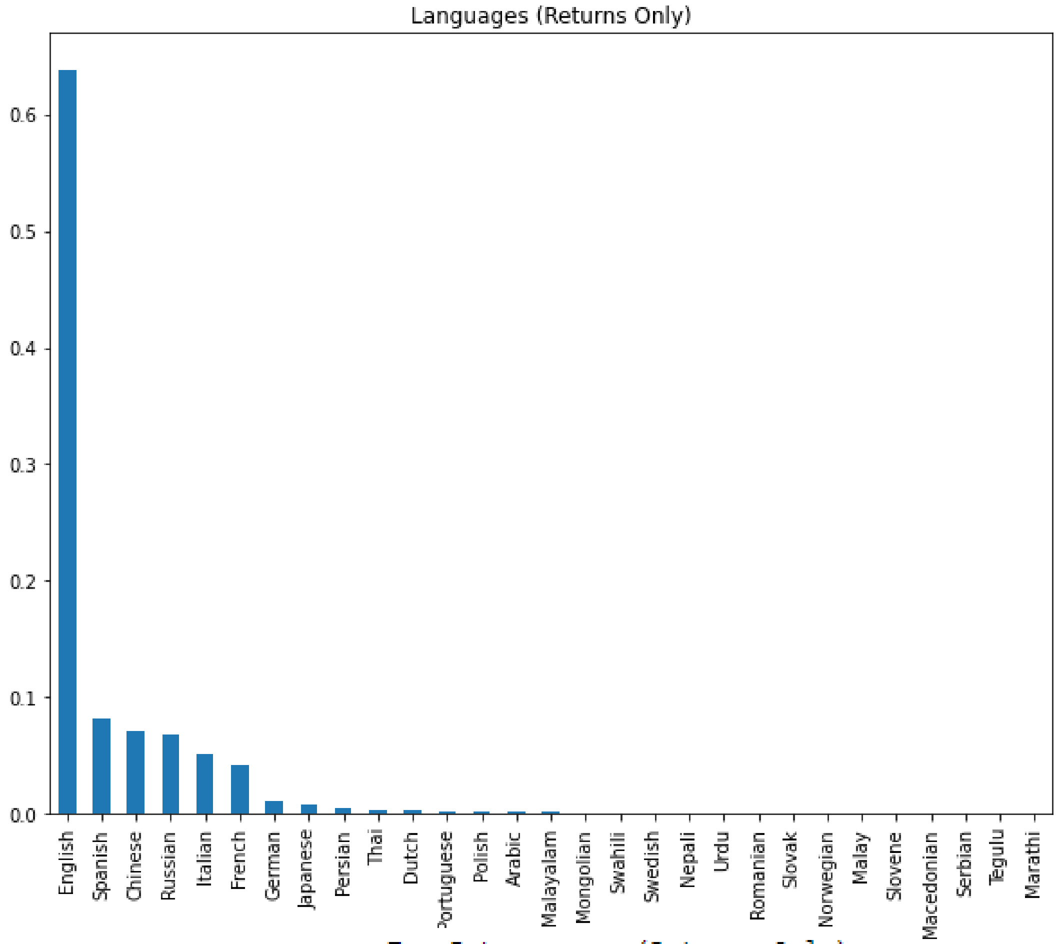
JP 6.0%

CN 5.9%

- Similar distributions in the 2 charts

Data Exploration - Language





Top 5 Languages

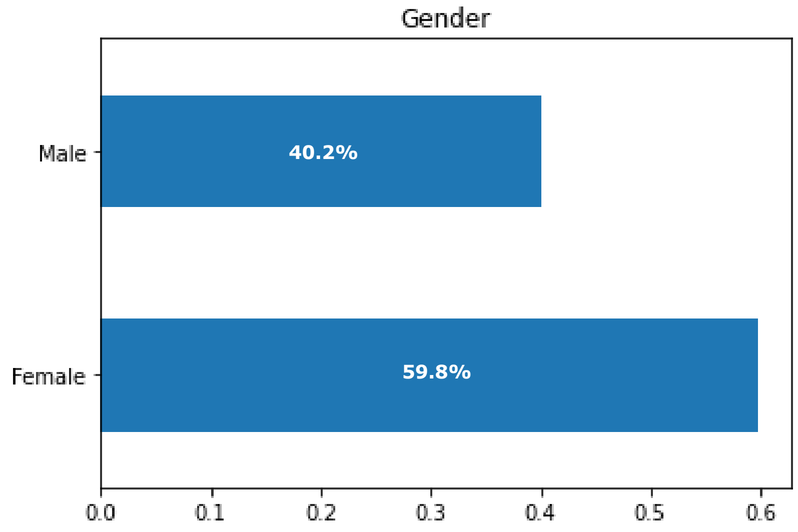
English	59.0%
Chinese	12.0%
Spanish	8.4%
Russian	7.1%
French	4.7%

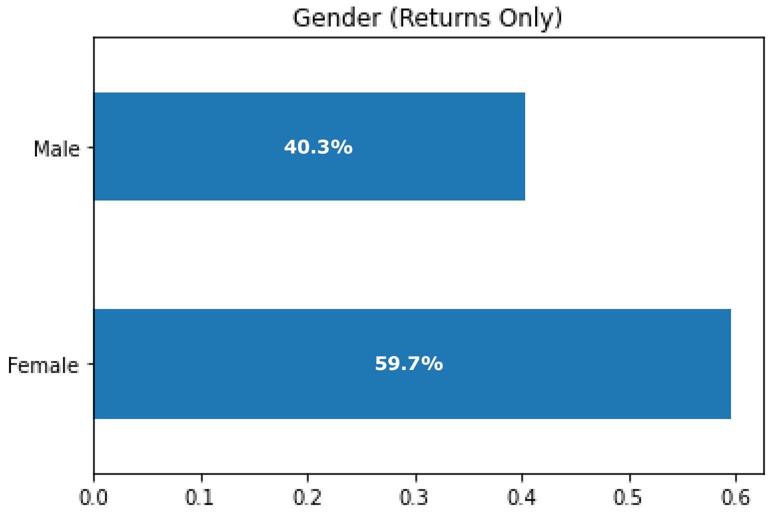
- Very similar distributions in the 2 charts

Top 5 Languages (Returns Only)

English	63.8%
Spanish	8.2%
Chinese	7.1%
Russian	6.8%
Italian	5.2%

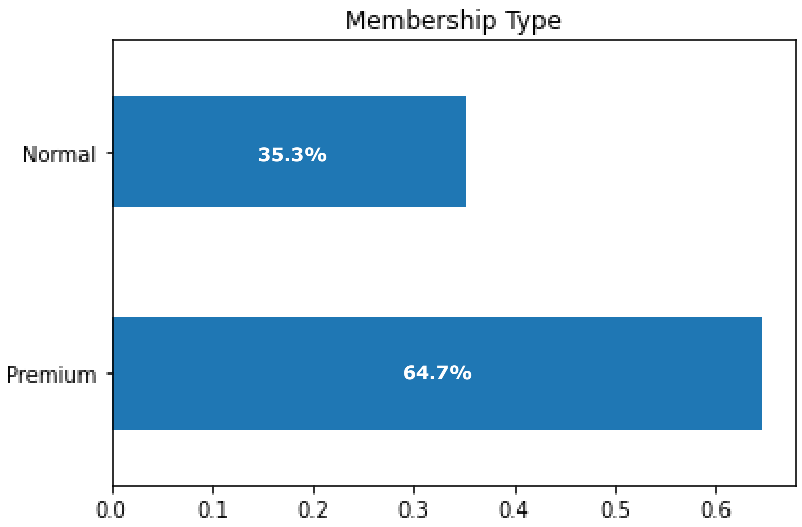
Data Exploration - Gender

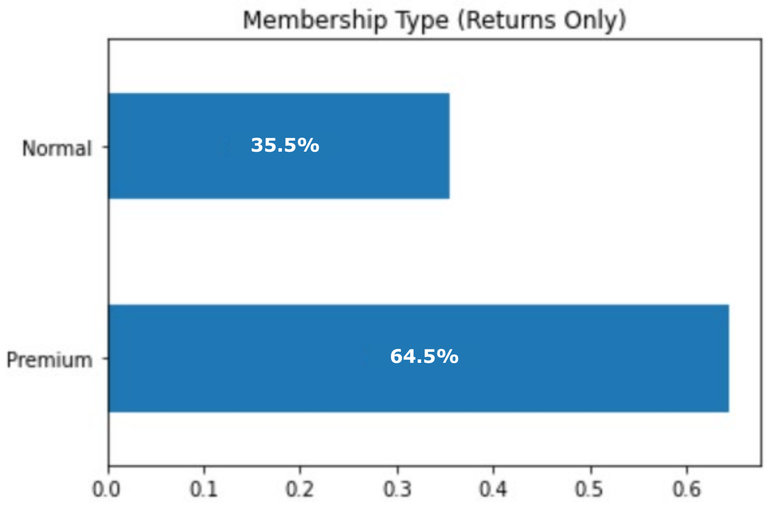




- Nearly identical split in the whole and subset

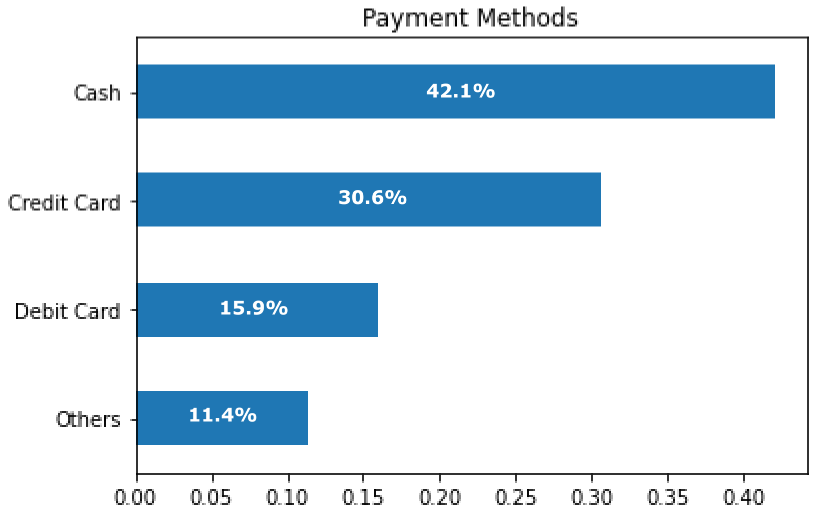
Data Exploration – Membership Type

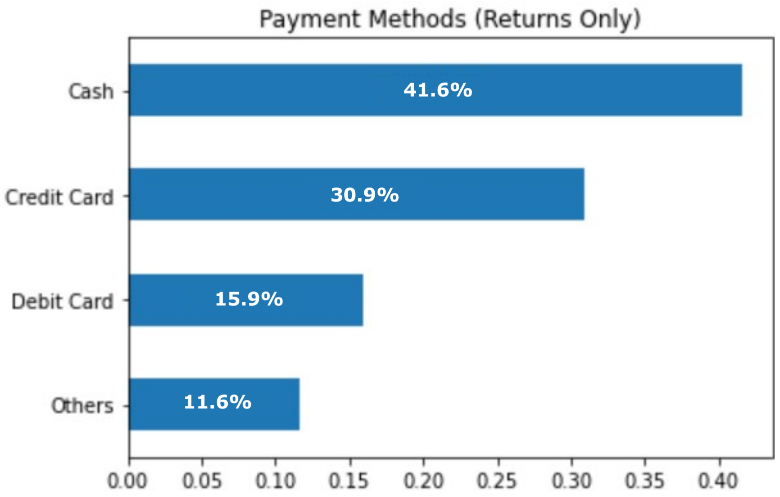




- Nearly identical split in the whole and subset

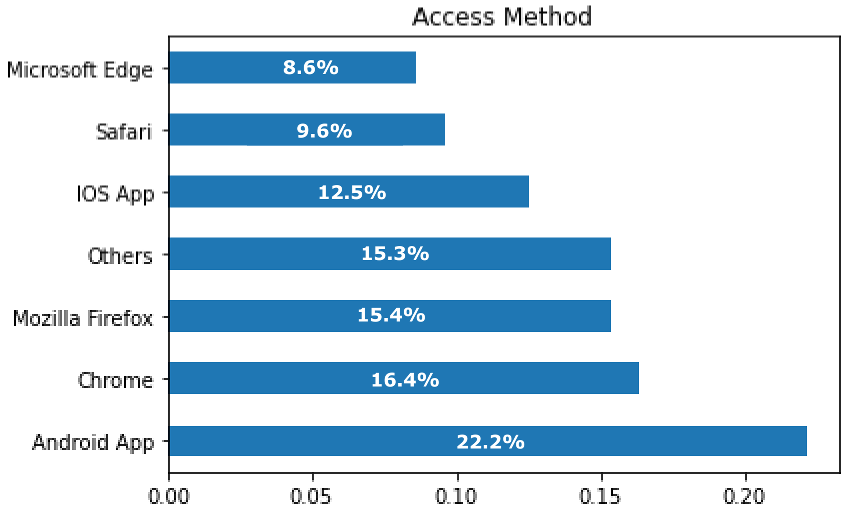
Data Exploration – Payment Methods

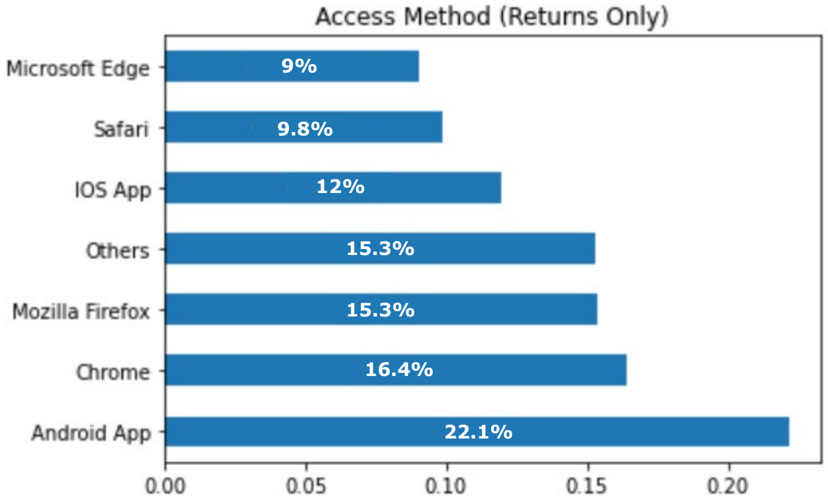




- Very similar splits in the whole and subset

Data Exploration – Access Method





- Nearly identical split in whole and subset

Methods

Imbalanced Data Problem

01

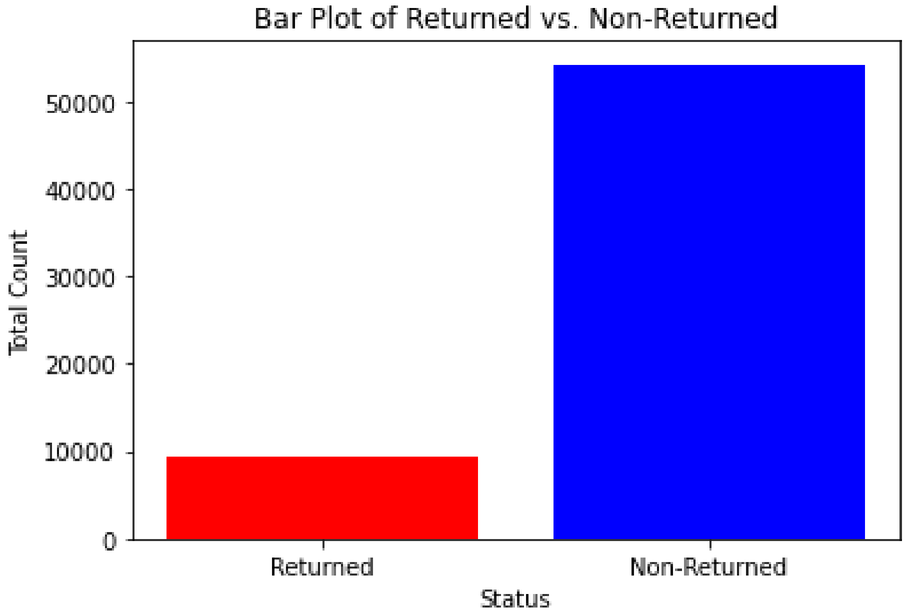
Biased Models
Possibility of bias toward the majority class, resulting in poor performance on the minority class

02

Difficulty in Model Evaluation
Metrics such as accuracy can be misleading as high accuracy can be achieved by predicting the majority class

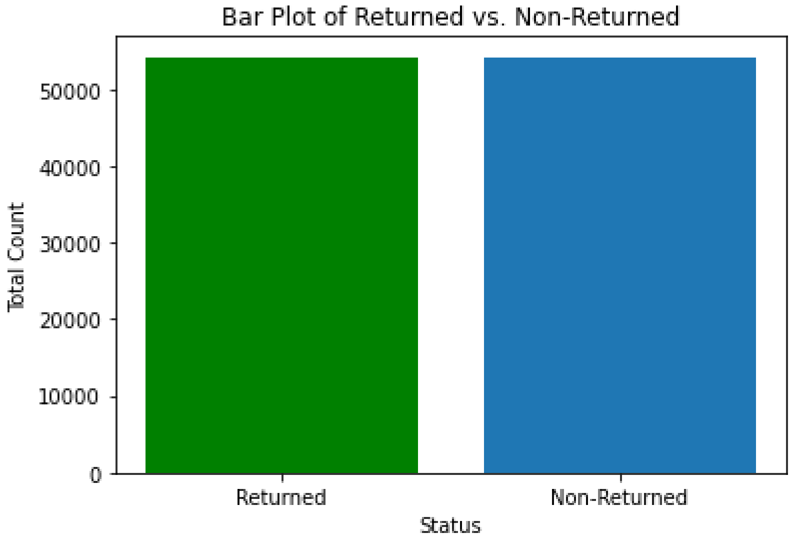
03

Feature Importance Bias
Model may focus more on features that help distinguish the majority class, neglecting features important for minority class prediction

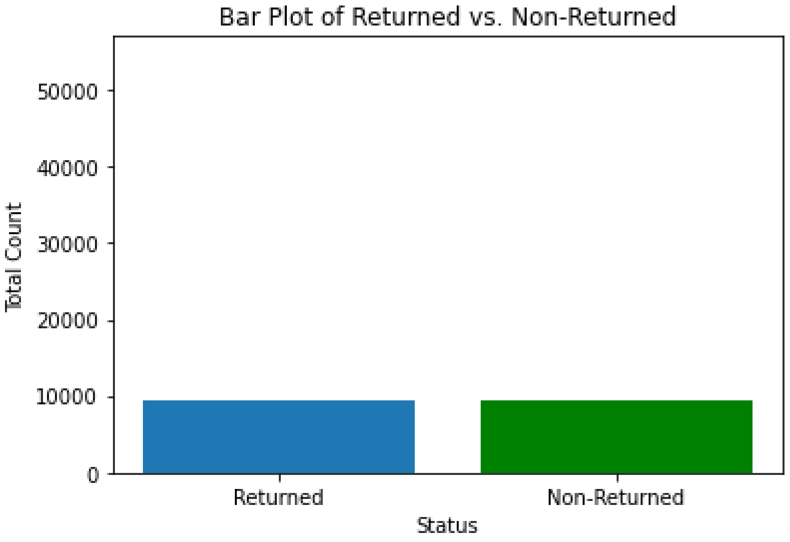


Solution: Oversampling/Under sampling Data

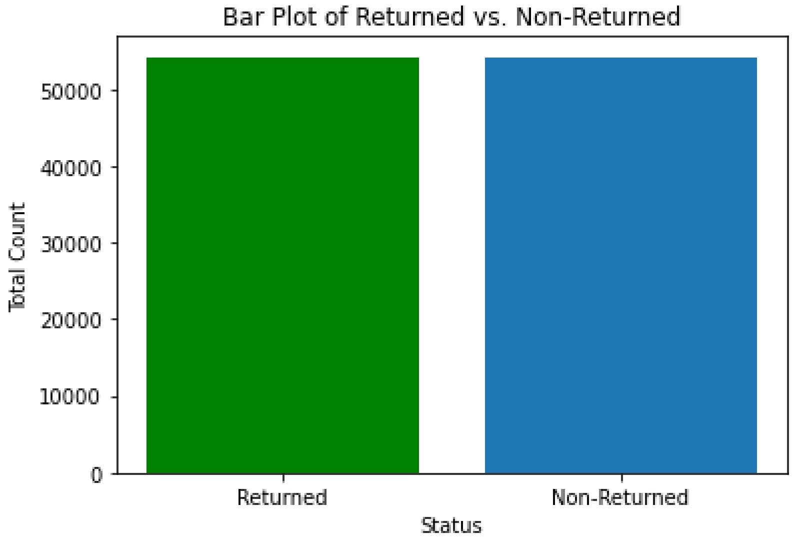
Oversampling



Under sampling



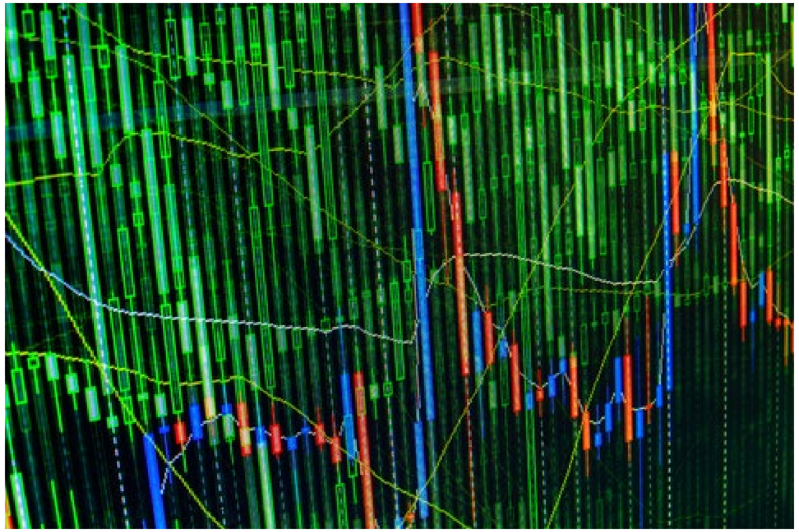
SMOTE sampling



- Oversampling and Under sampling was essential in fixing the data imbalance issue
- Both classes had equal distributions of data

Logistic Regression

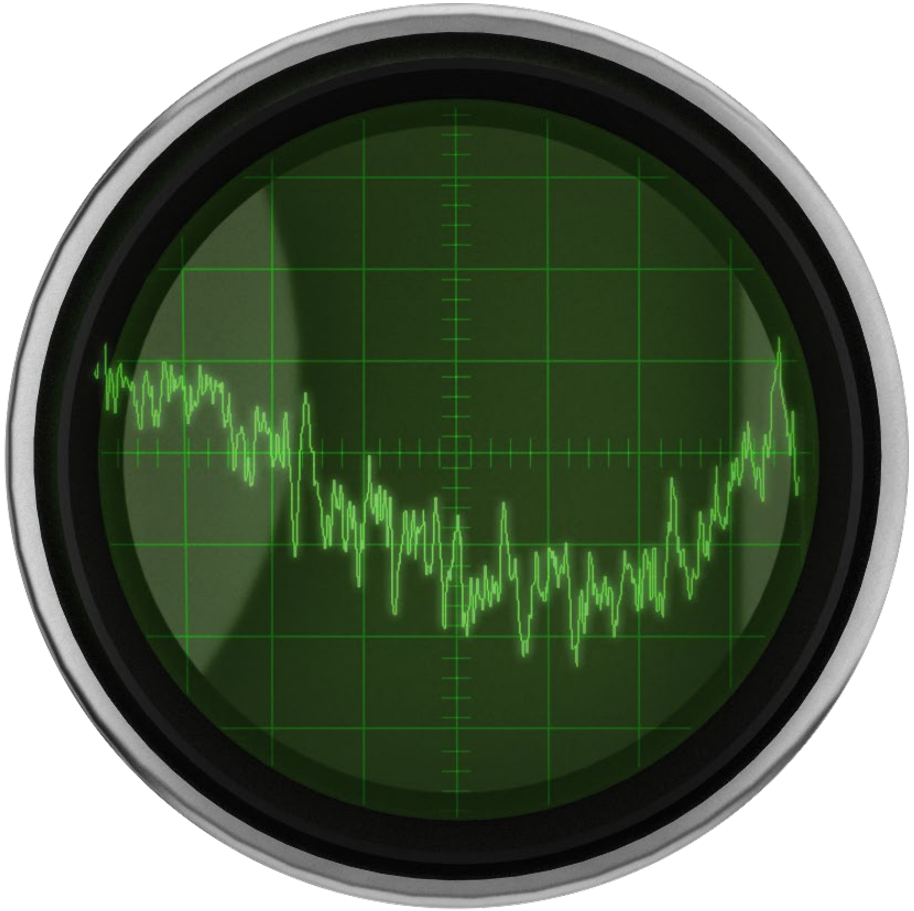
- Statistical model used for binary classification tasks where target variable is categorical and has only two possible outcomes
- Estimates the probability that a given input belongs to one of the classes by fitting a logistic function to the observed data



Random Forest

- Random Forest is a learning algorithm used for classification and regression tasks in machine learning
- Belongs to the family of tree-based methods known for its high predictive accuracy
- Used for binary classification tasks where the target variable is categorical with two possible outcomes capturing complex relationships and provide robust predictions

Metrics/Results



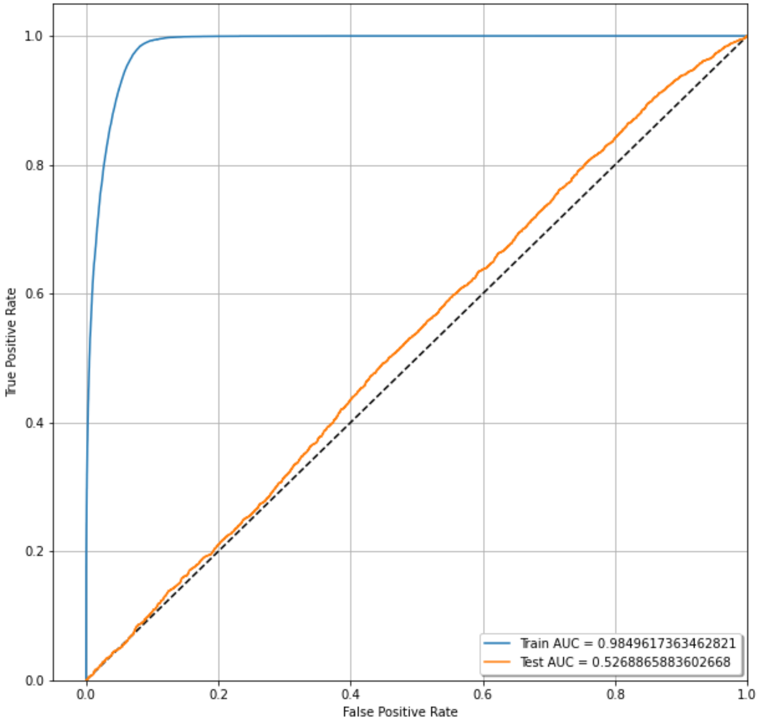
Accuracy Metrics

Training	**Logistic regression**	**LogReg GridSearch**	**Random Forest**	**RanFor GridSearch**
Oversampling	53.95%	53.86%	95.87%	94.95%
Undersampling	54.19%	54.39%	96.37%	78.18%
SMOTE	77.37%	77.90%	94.68%	89.54%

Testing	**Logistic regression**	**LogReg GridSearch**	**Random Forest**	**RanFor GridSearch**
Oversampling	40.80%	40.70%	73.95%	72.63%
Undersampling	39.66%	39.60%	51.90%	46.54%
SMOTE	74.03%	75.62%	70.79%	66.20%

Data Visualization

ROC Curve



AUC Results

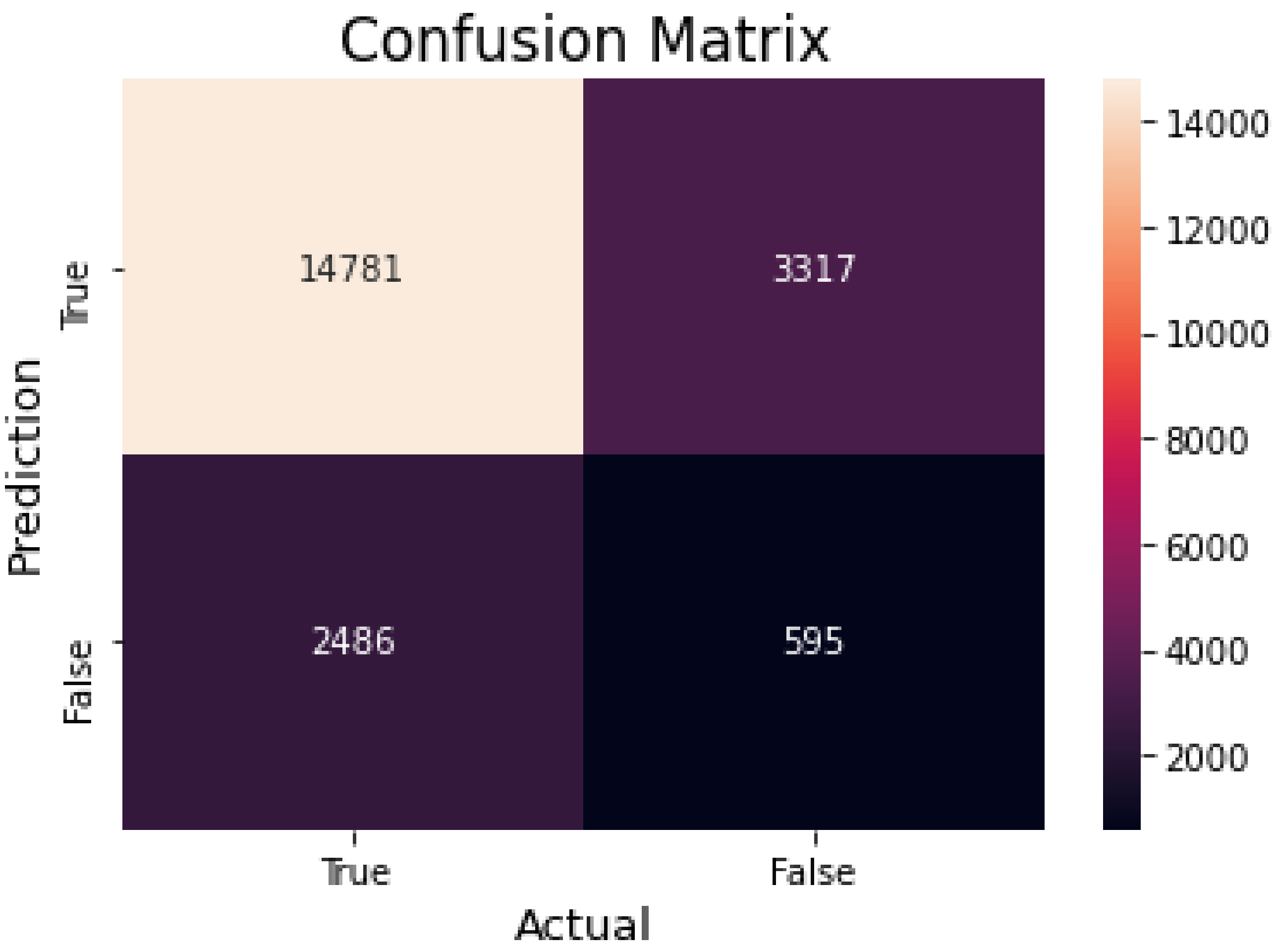
0.985 AUC

Training

0.527 AUC

Testing

Data Visualization



Precision and Recall Metrics

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)}$$

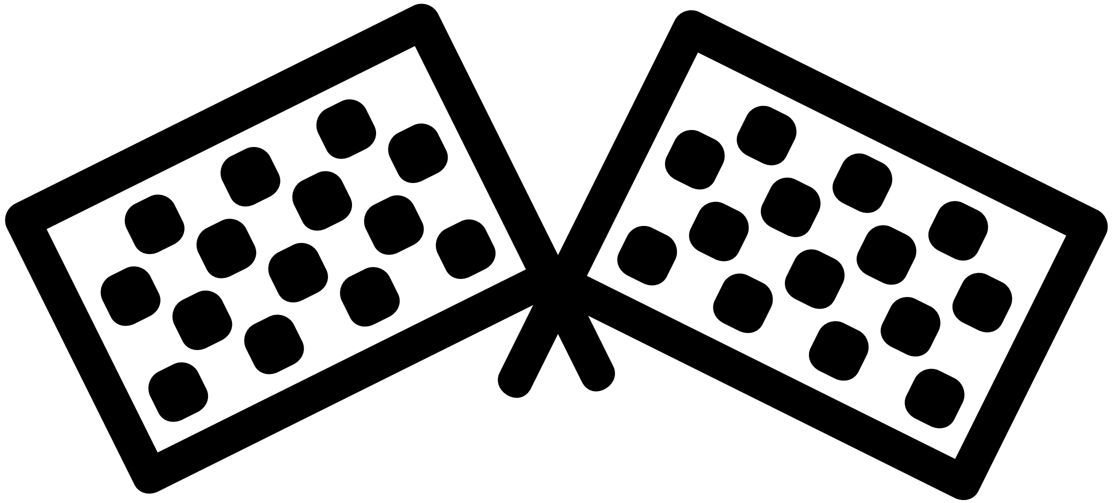
$$Recall\ = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

Precision and Recall Metrics

- The precision and recall is higher when predicting non-returned rather than predicting returned

	precision	recall	f1-score	support
0	0.86	0.82	0.84	18098
1	0.15	0.19	0.17	3081
accuracy			0.73	21179
macro avg	0.50	0.50	0.50	21179
weighted avg	0.75	0.73	0.74	21179

Conclusion



Conclusion

Key Takeaways

- **Model still is dealing with imbalanced data problems**
- **The distribution of columns in the whole dataset and just for returned items only are about the same**
- **Random Forest without the hyperparameters had the highest accuracy**

Next Steps

More data

Richer reviews

Fit a proper neural network to the dataset

Add additional data collection for returns (ex. 5 reasons why the items were returned)

Implement a more robust returns process

Deloitte.

Questions?