Bank Churners

MAHA ALADWANI
Mohammed AlShehri
MAHA ALHARQAN
YAZEED ALHARTHI

Junior Data Scientists
at SDAIA Academy BOOTCAMP T5

Today's Topics

OUR DISCUSSION FLOW

- Introduction
- About Dataset
- Challenges
- Questions
- Algorithms
- Conclusions
- Tools

Introduction

The main purpose of this project is to classify existing and attrited customers by using classification models.
To know them so to provides them with better services, and turn customer decisions in the opposite direction.

About Dataset

- This dataset can be found at **Kaggle**.
- It has 10127 rows and 23 columns

Challenges

Dealing with null value

Dealing with imbalanced data

Choosing the right hyperparameter

Questions

01

What is the most Income Category based on Attrition Flag?

02

Is the Education Level affect the Income Category?

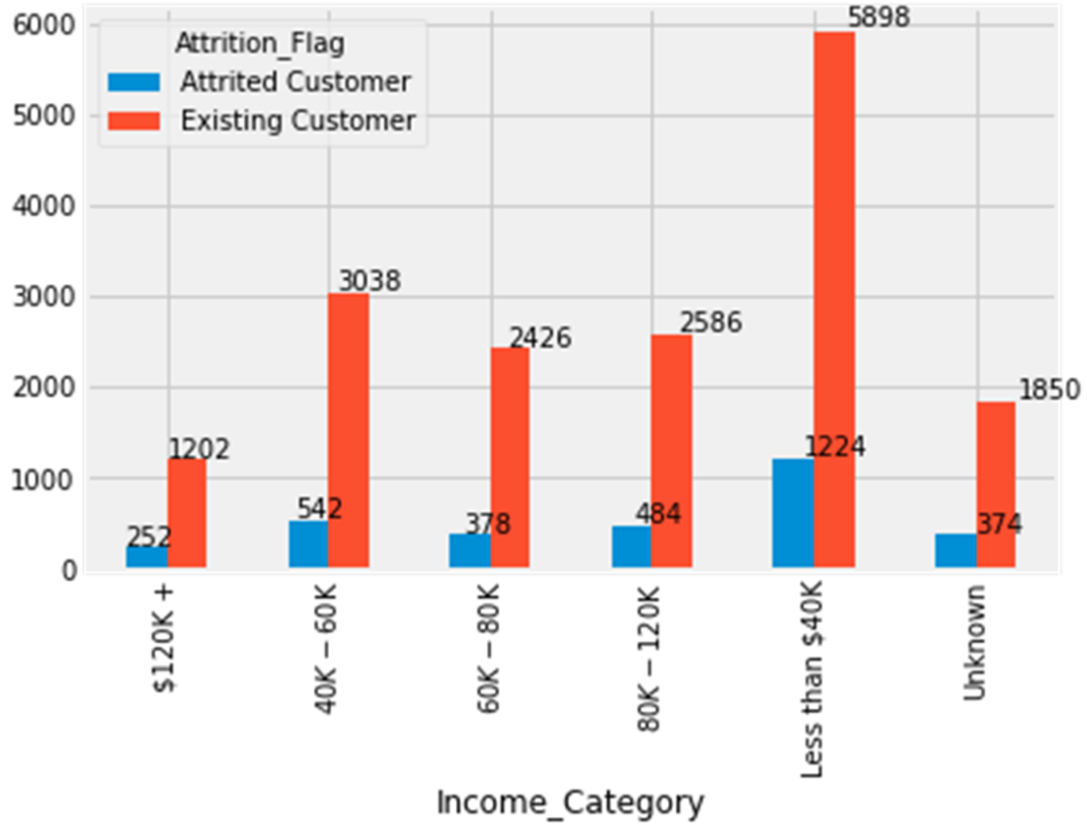
03

Does Customer Age influence the Marital Status?

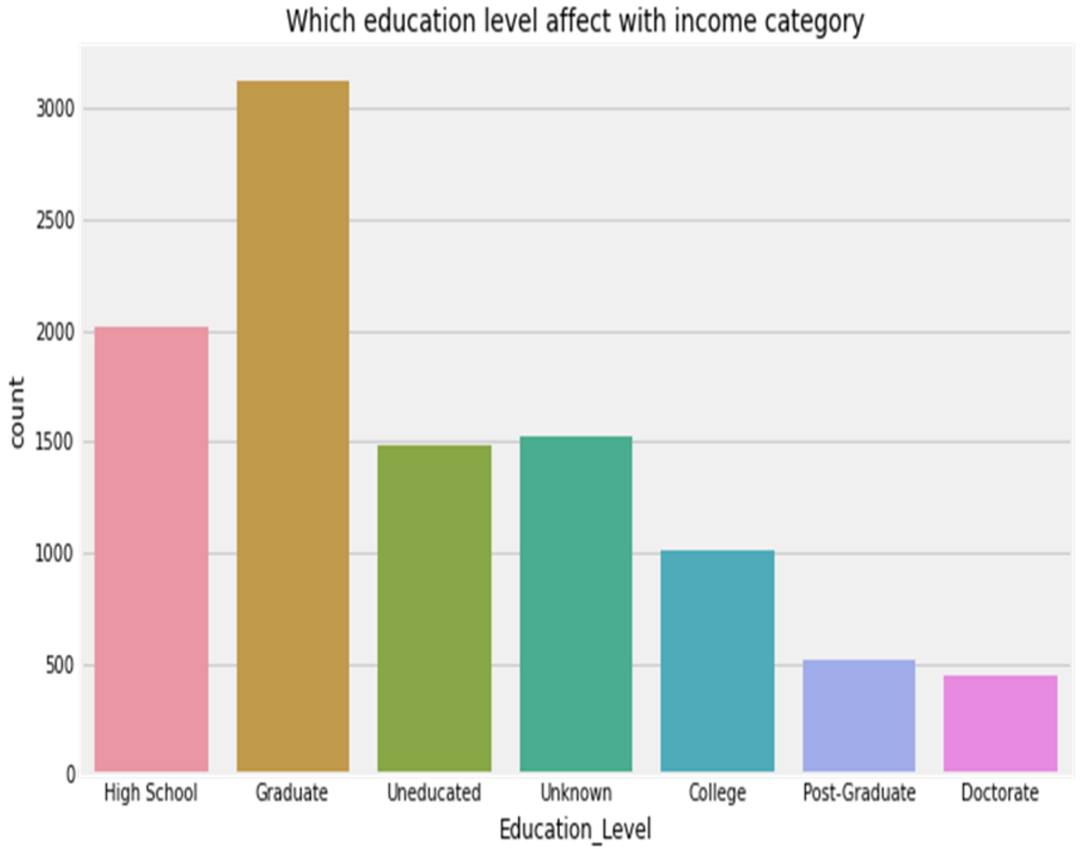
04

Which card category has been used the most?

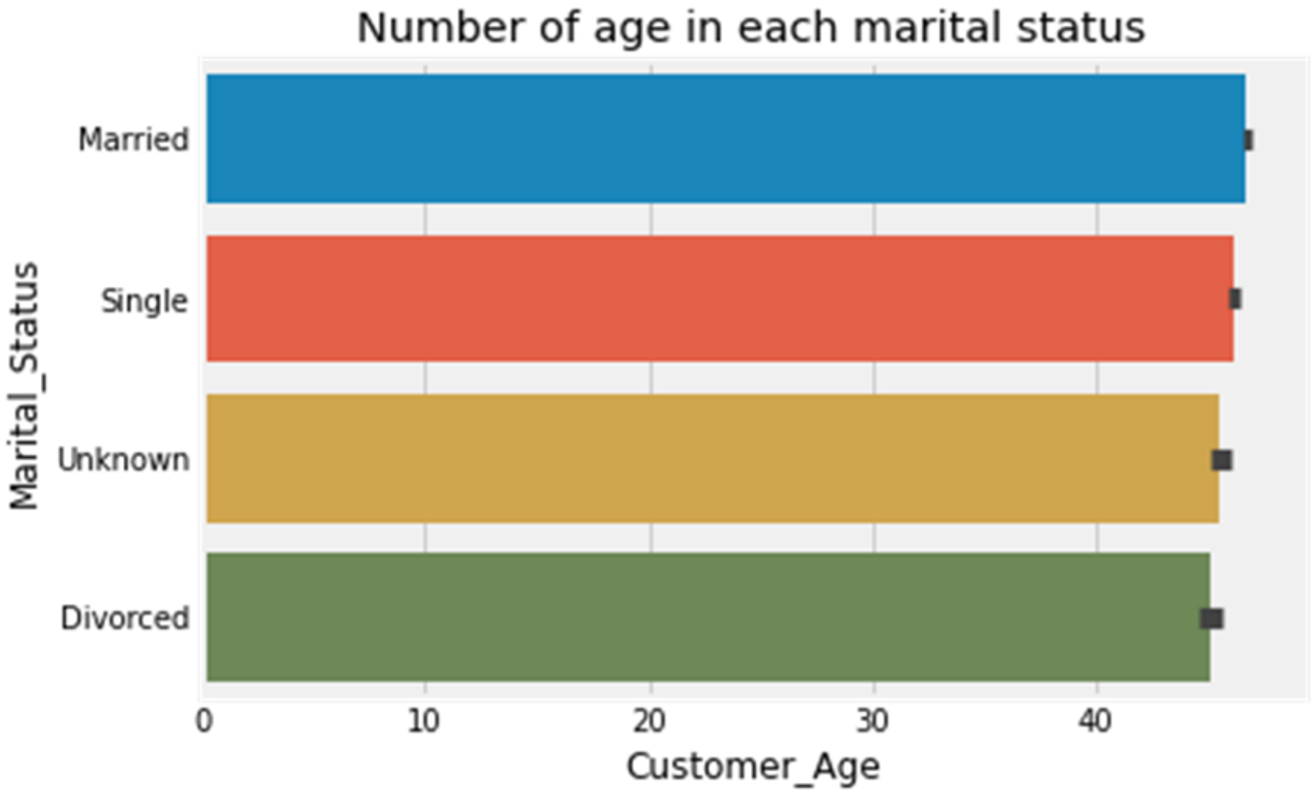
What is the most Income Category based on Attrition Flag?



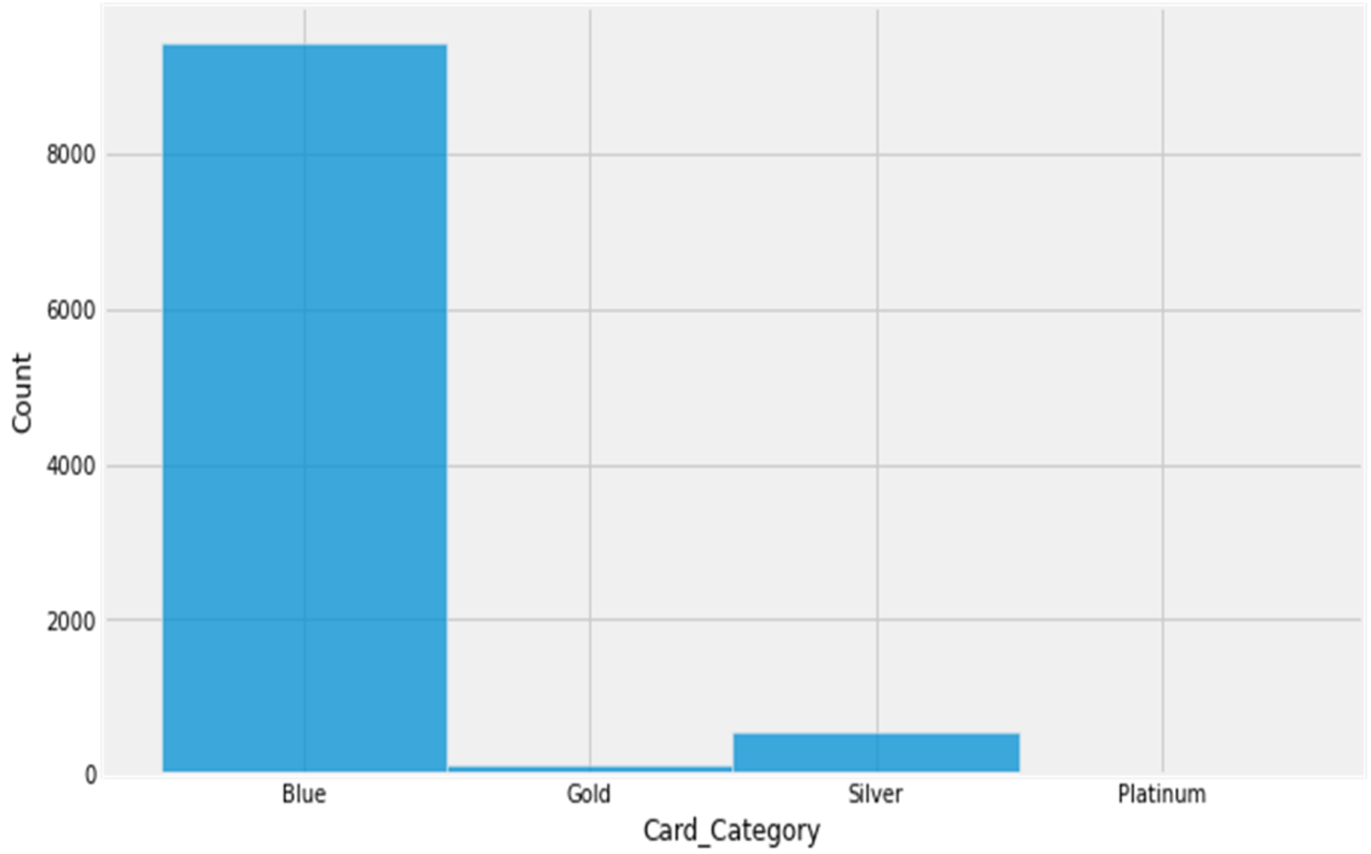
Is the Education Level affect the Income Category?



Does Customer Age influence the Marital Status?



Which card category has been used the most?



Algorithms

- K-Nearest Neighbor
- Logistic Regression
- Decision Tree
- Random Forest
- Extreme Gradient Boost
- Light Gradient Boost
- Stacking

The prediction results for the attrited customers

After dealing with the imbalanced data

MODEL	Accuracy	Recall	Precision	F1-Score
K-Nearest Neighbor	91.79%	96.62%	88.24%	92%
Logistic Regression	93.67%	91.61%	95.68%	94%
Decision Tree	94.17%	95.45%	93.17%	94%
Random Forest	95.0%	95.69%	94.47%	95%
Extreme Gradient Boost	98.29%	98.36%	98.25%	98%
Light Gradient Boost	98.29%	98.42%	98.19%	98%

Comparison of the Light Gradient Boost model

Before and after dealing with the imbalanced data

LGB MODEL	Accuracy	Recall	Precision	F1-Score
Before	97.03%	86.46%	94.61%	92%
After	98.29%	98.42%	98.19%	98%

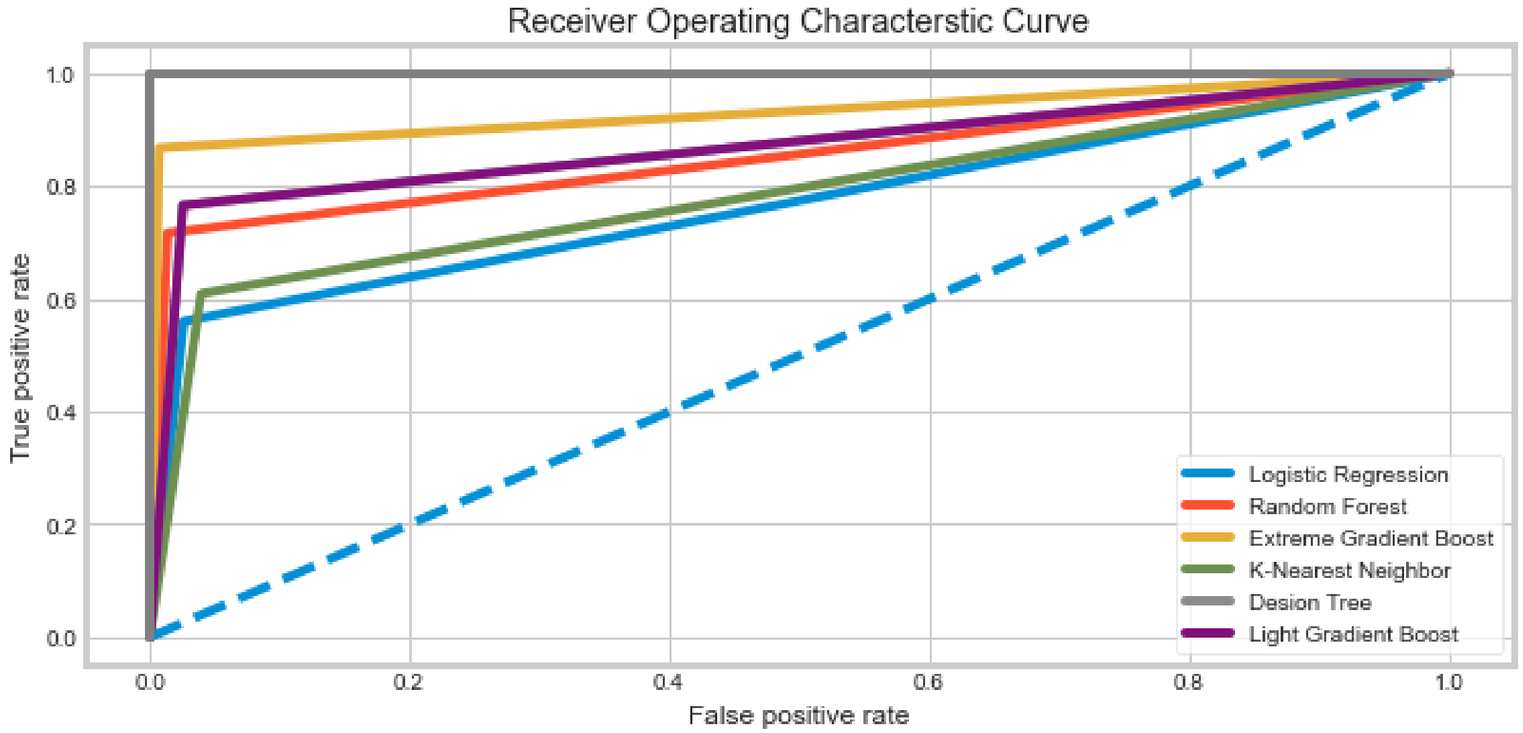
Comparison of the lgbm with the staking results

After dealing with the imbalanced data

MODEL	Accuracy	Recall	Precision	F1-Score
The stacking	98.17%	98.31%	98.08%	98%
Light Gradient Boost	98.29%	98.42%	98.19%	98%

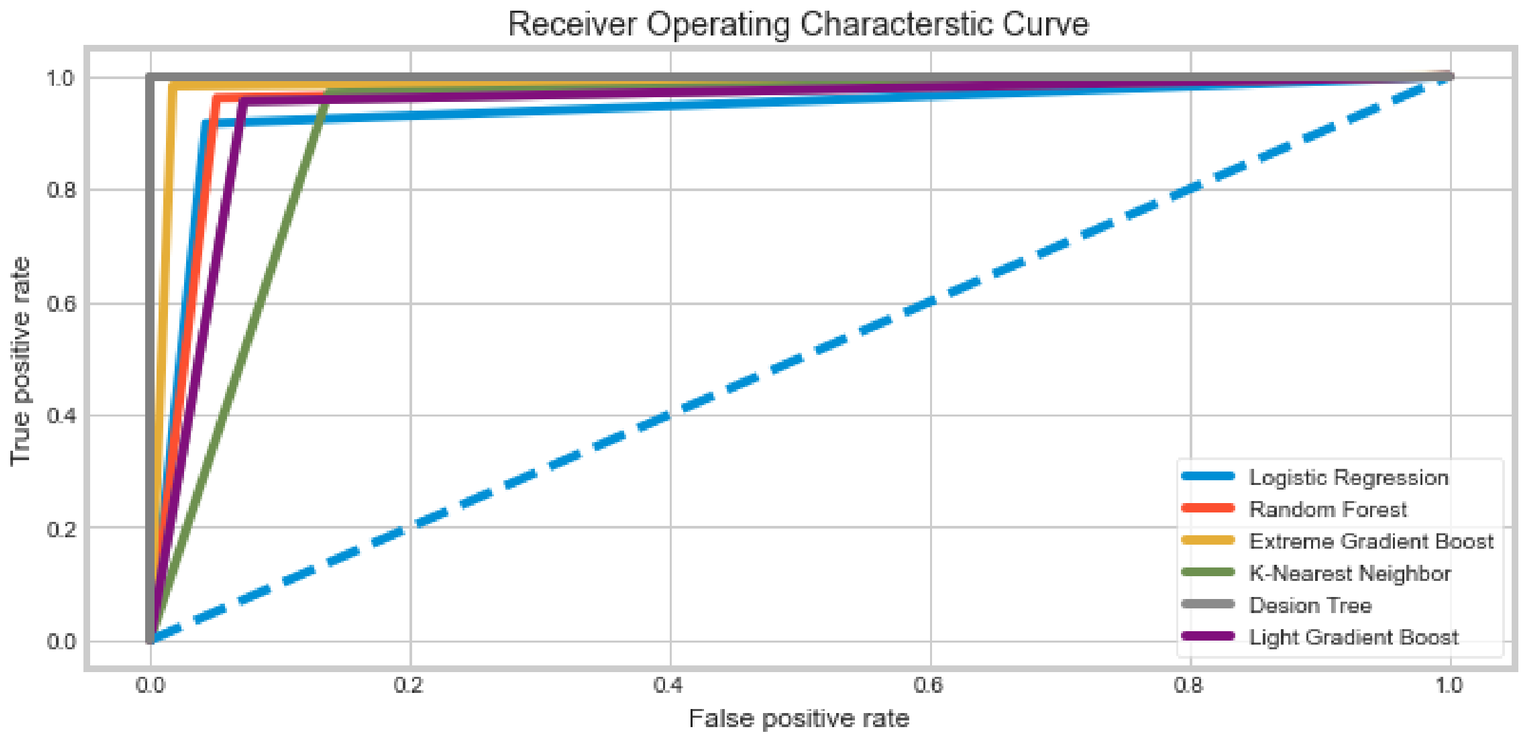
Receiver operating characteristic curve

With the imbalanced data



Receiver operating characteristic curve

After dealing with the imbalanced data



Tools

- Jupyter
- Python
- Pandas
- Numpy
- Zoom
- Seaborn
- Scikit Learn

Conclusion

- We found that the Light Gradient Boost model has the best results among the other models.
- We found that the lowest rate of recall for the models is Logistic Regression.
- Most of the customers had a blue card, so we need to market each card to build diverse categories of customers
- We can use these models on bank websites to predict the number of users
- We think that it is not necessary to use the stacking because it did not have much impact.

Thank You!