ta2menak

Supervised by

DR.Ihab El-Khodary
TA. Mohie

Prepared by

Mayar Ahmed
Mona Abdelaziz
Habiba Abdelkareem
Dina Ahmed

Table of content

Introduction

Car insurance plays a pivotal role in safeguarding individuals and their vehicles in today's dynamic and ever-evolving world. As the number of vehicles on the roads continues to rise, the imperative for comprehensive car insurance becomes increasingly evident. This financial protection not only ensures the well-being of the vehicle owner but also contributes to the overall safety and security on the roads.

Problem statement

A lot of problems can face car insurance companies if they assets their risk in wrong way such as.
Insurance companies calculate the risk in a rigid manner based on the status of the car only, without taking into considiration the car driver’s risk ,which will lead to Financial instability and Increased premiums.

Objectives

1

Risk Assessment

To assess risk for car insurance customers to minimize it as much as possible.

2

Risk Score Classification

To accept or reject making new insurance policies for customers depending on their risk levels (high , medium and low).

3

Claim Prediction Model

To predict if the customer will claim the insurance premium within 6 months or not.

Data

At first we try to find data in Eygpt but the car insurance companies refused to give us a real data but just a limited reports.

We searched on websites for real data in Eygpt but we find nothing. So our biggest problem was the data limitation which was:

- Difficulty in finding data
- Limited Data Availability
- Data Quality Issues

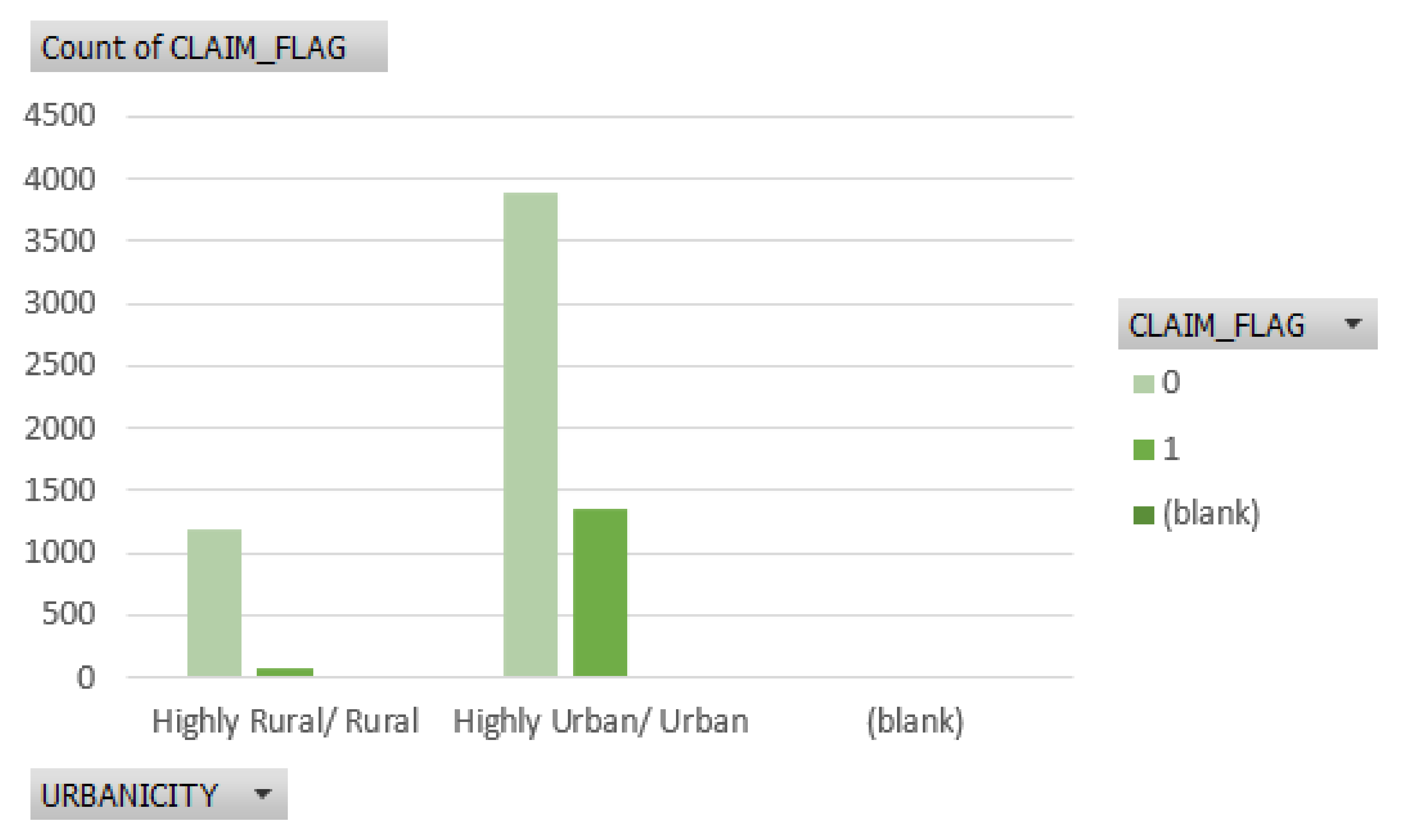
Data cleaning

- we start with dropping some columns “red car,birth ” .
- we fill the blank cells of the columns with the average “income, blue book, home value”.

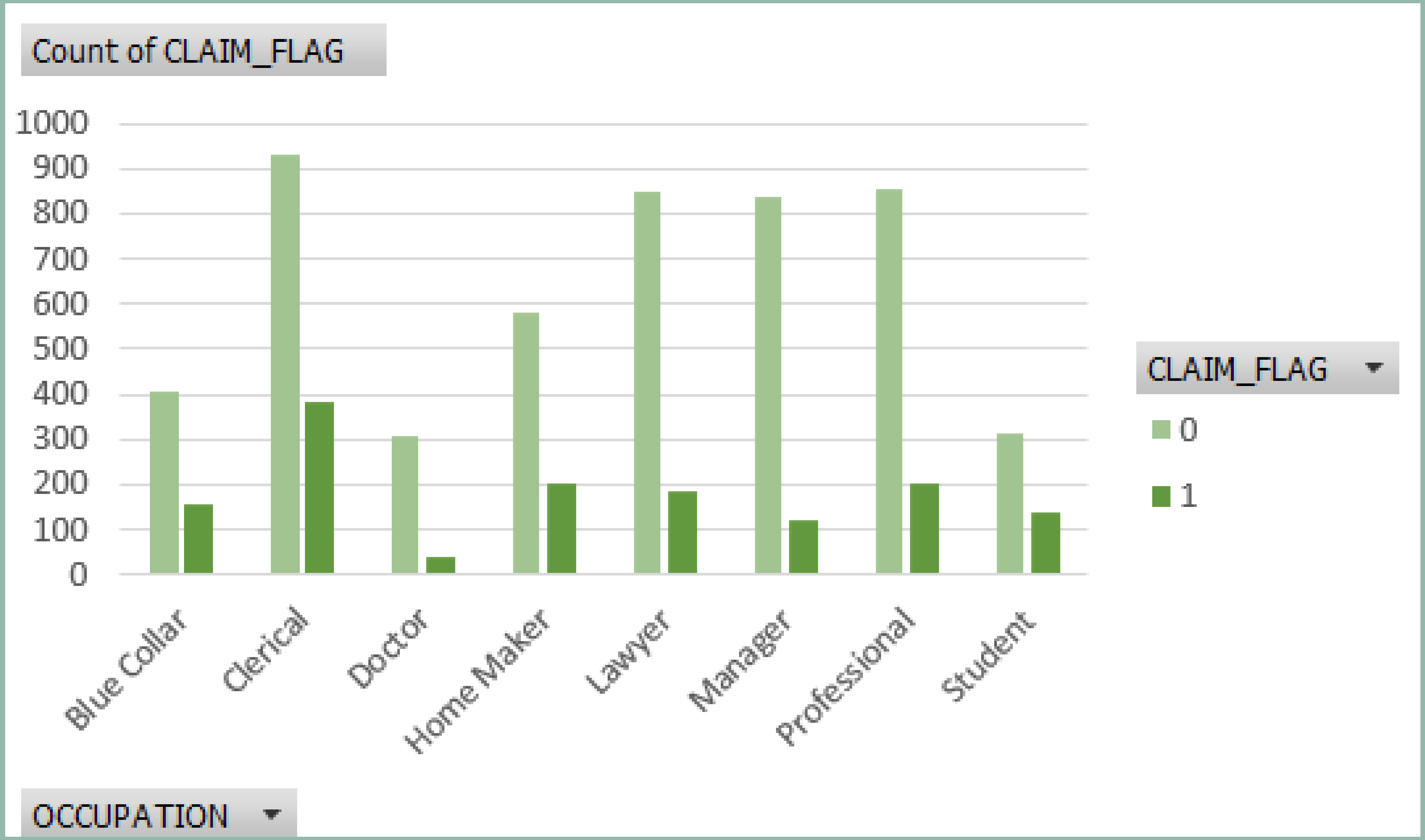
The effective columns in calculating the risk:

- KIDSDRIV
- AGE
- PARENT1
- EDUCATION
- OCCUPATION
- TIMEINFORCE
- TRAVTIME
- CLM_AMT
- OLDCLAIM
- CLM_FREQ
- REVOKED
- URBANICITY

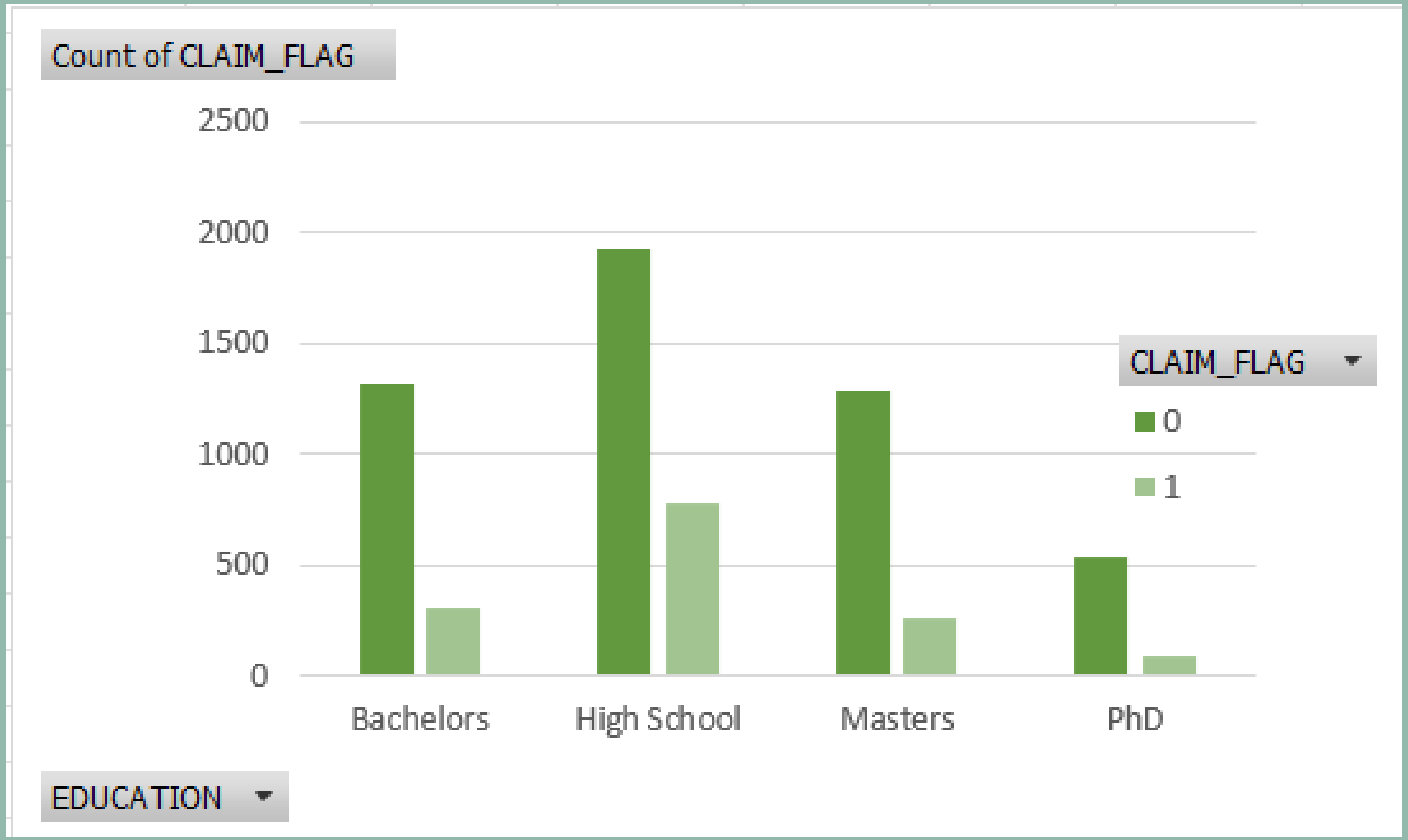
sum of the data analysis



sum of the data analysis



sum of the data analysis



models we used

why logistic regression as base model

Simplicity and interpertability

Logistic regression is easy to understand and implement. The model coefficients provide insights into how each feature influences the outcome, making it a good choice for building interpretable models.

Efficiency

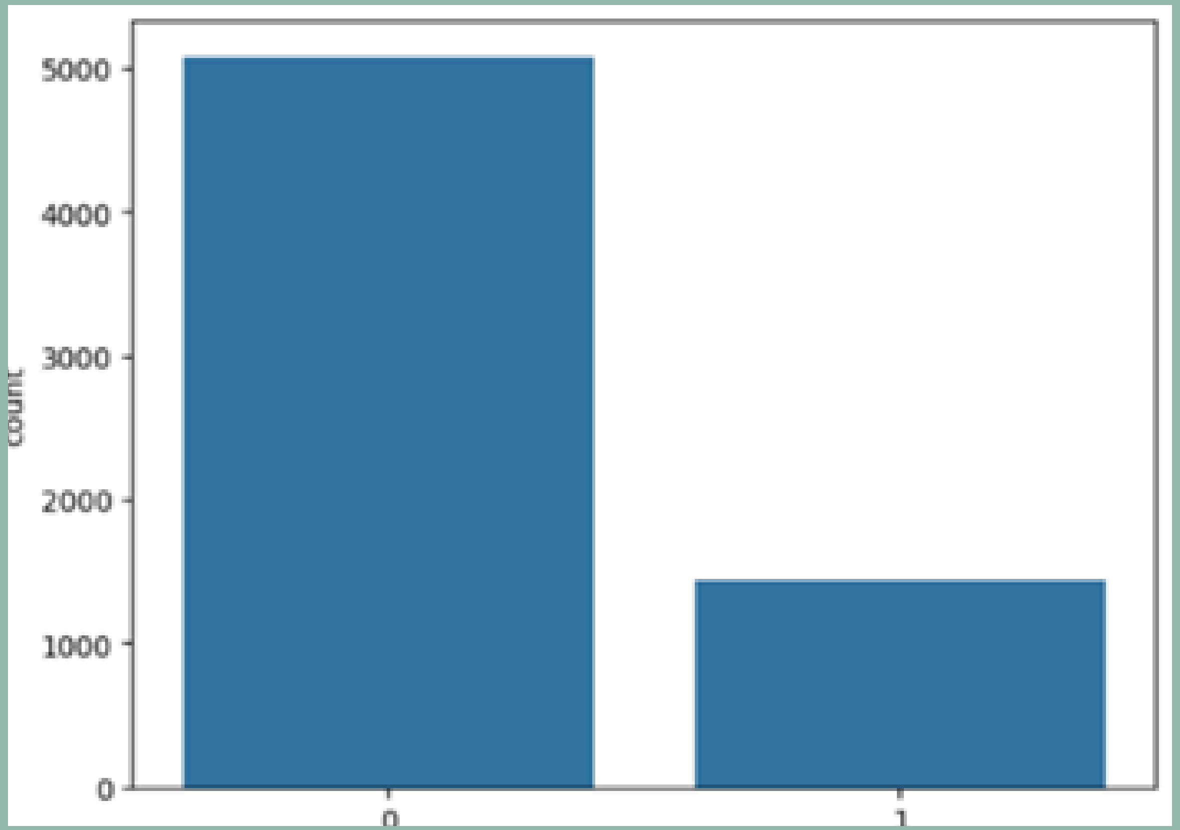
Training a logistic regression model is computationally efficient, especially compared to more complex algorithms like neural networks. This makes it ideal for large datasets.

Probabilistic Outputs

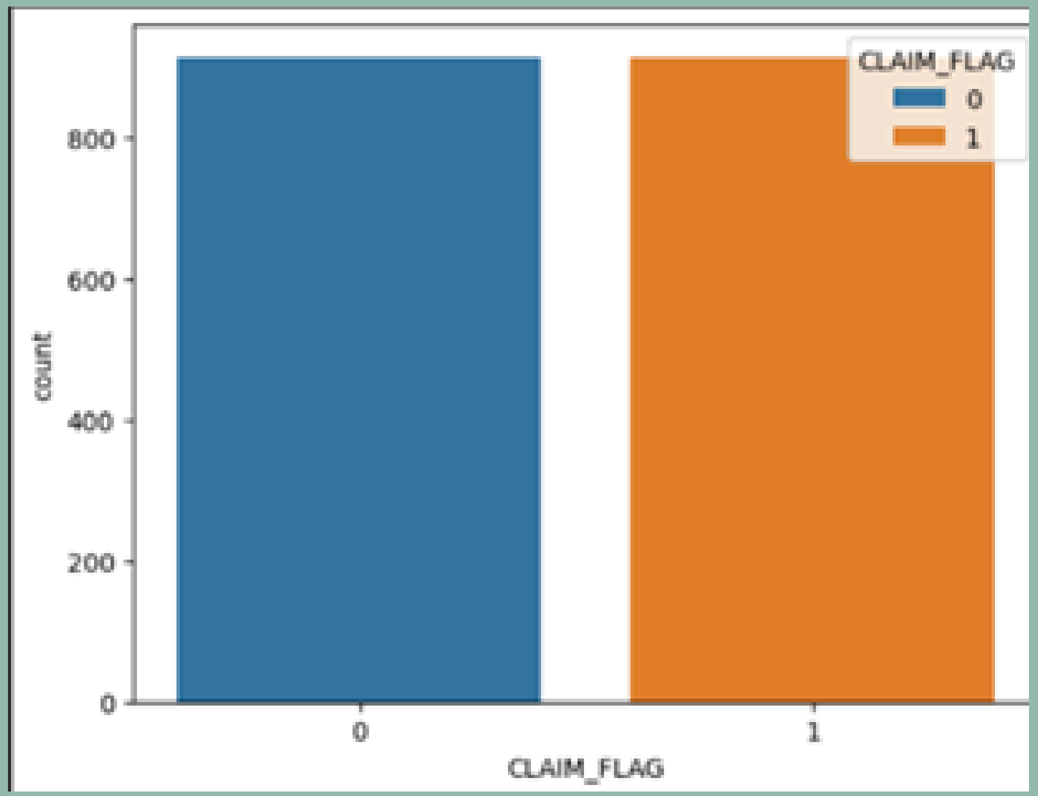
Logistic regression doesn't just predict a class label (e.g., spam/not spam), it also outputs the probability of belonging to each class. This provides valuable confidence scores for predictions

challenge we faced in logistic regression and how we overcome it?

target column in data was imbalance so we used undersampling to handle it



target column before undersampling



target column after undersampling

types of undersampling:

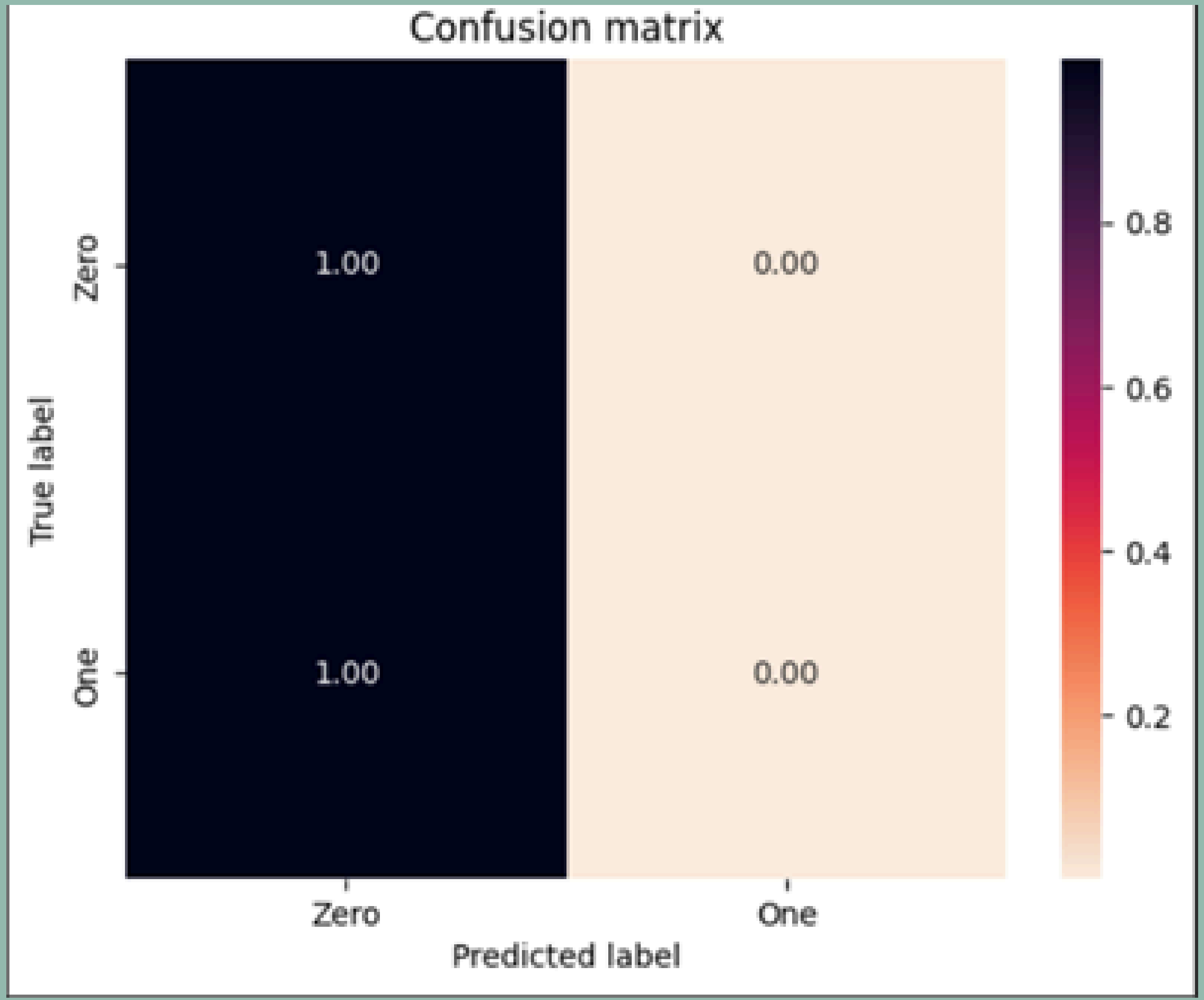
1

undersampling based on majority

undersampling based on minority

2

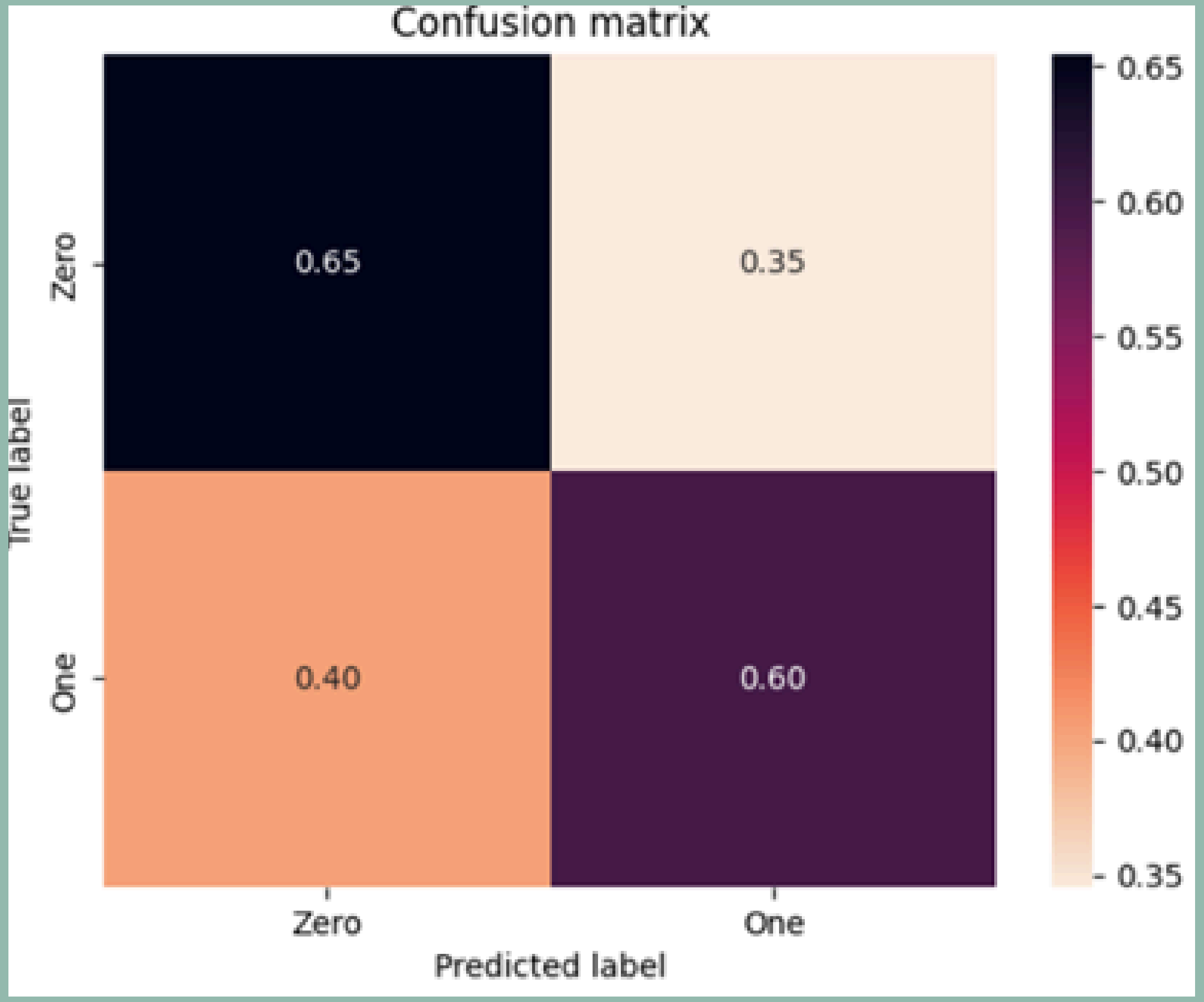
undersampling based on minority



	precision	recall	f1-score	support
0	0.78	1.00	0.88	1532
1	0.50	0.00	0.01	423
accuracy			0.78	1955
macro avg	0.64	0.50	0.44	1955
weighted avg	0.72	0.78	0.69	1955

(confusion matrix in minority undersampling)
Accuracy 75.15%

undersampling based on majority



	precision	recall	f1-score	support
0	0.62	0.65	0.64	402
1	0.63	0.60	0.61	402
accuracy			0.63	804
macro avg	0.63	0.63	0.63	804
weighted avg	0.63	0.63	0.63	804

(confusion matrix in majority undersampling)
Accuracy 78.22%

Enhancement technique

Car incurance	Boosting	Bagging
Personal behavioral data	Variability and Complexity of Behavioral Data	Reducing Overfitting
Importance of Accuracy	Importance of Accuracy	Handling High Variance
Handling Outliers and Rare Events	Handling Outliers and Rare Events	Robustness to Noise

Why XGBoost ?

XGBoosting vs LightGBM

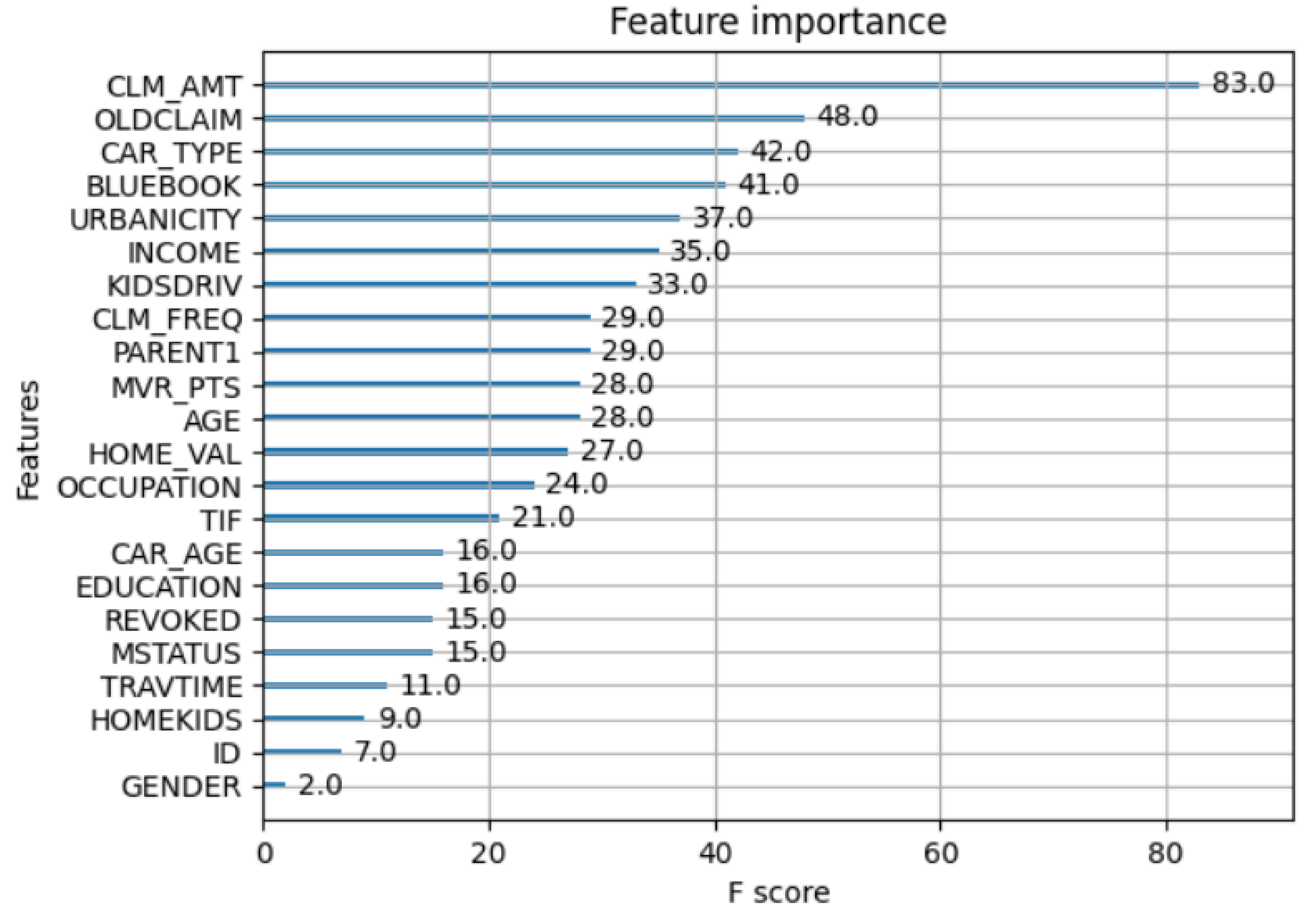
XGBoost generally offers better regularization and handles sparse data more effectively, although LightGBM can be faster in certain scenarios

XGBoosting vs AdaBoost

XGBoost's gradient boosting framework with advanced regularization and tree pruning makes it more robust and accurate than AdaBoost.

Feature engineering in XGBoost ?

the relative contribution of each feature to the prediction model. most influential in predicting car insurance risks.



Classification in XGBoost ?

From zero to 0.33 low
0.33 to 0.66 medium
bigger than 0.66 High
even if meduim classification is a critical situation based on decision maker

	ID	Prediction	Probability	Risk
0	453194620	0	0.164559	Low
1	794901879	0	0.482494	Medium
2	345352186	0	0.166901	Low
3	582345812	0	0.145945	Low
4	197012324	0	0.173631	Low
...	...	...	...	...
2601	918387800	0	0.434330	Medium
2602	554209949	0	0.161463	Low
2603	343314151	0	0.474839	Medium
2604	433559105	0	0.179718	Low
2605	413007803	0	0.151703	Low

XGBoost results

XGBoost increase the accuracy of the base model beacouse it is train several time and take the average of it at the end.

```
Accuracy:0.8038
```

conclusion

xgboost predicts with greater accuracy the probability that a person is more risky or not than logistics regression.

It can handle complex data better than logistics.

XGBoost algorithms are more flexible in dealing with different data sets, even those that are large in size or contain noise or missing data than logistics regression.

+

This increases the focus on specific categories of clients or specific types of risks.

This improves the model results Therefore, it improves accuracy (78.2% for logistics regression and 80.3% for XGboost)

Future Work

Enhanced Data Collection
Incorporate Advanced Features
Deployment and Integration
Continuous Monitoring and Updating

Time line

October
Define the idea
and Search about real
data

December
Problem formulation

November
Comparison of suitable
models for data

January/ February
Data preparation

Time line

March
searching for model

May
We chose logistic
regression and xgboost
,started building models

April
Comparison of suitable
models for data

June /July
Training models and
documentation
The project

Thanks!