CAR INSURANCE

PREDICTING RISK OF
CONFIRMED USE OF
INSURANCE WITHIN 6 MONTHS.

MEMBER

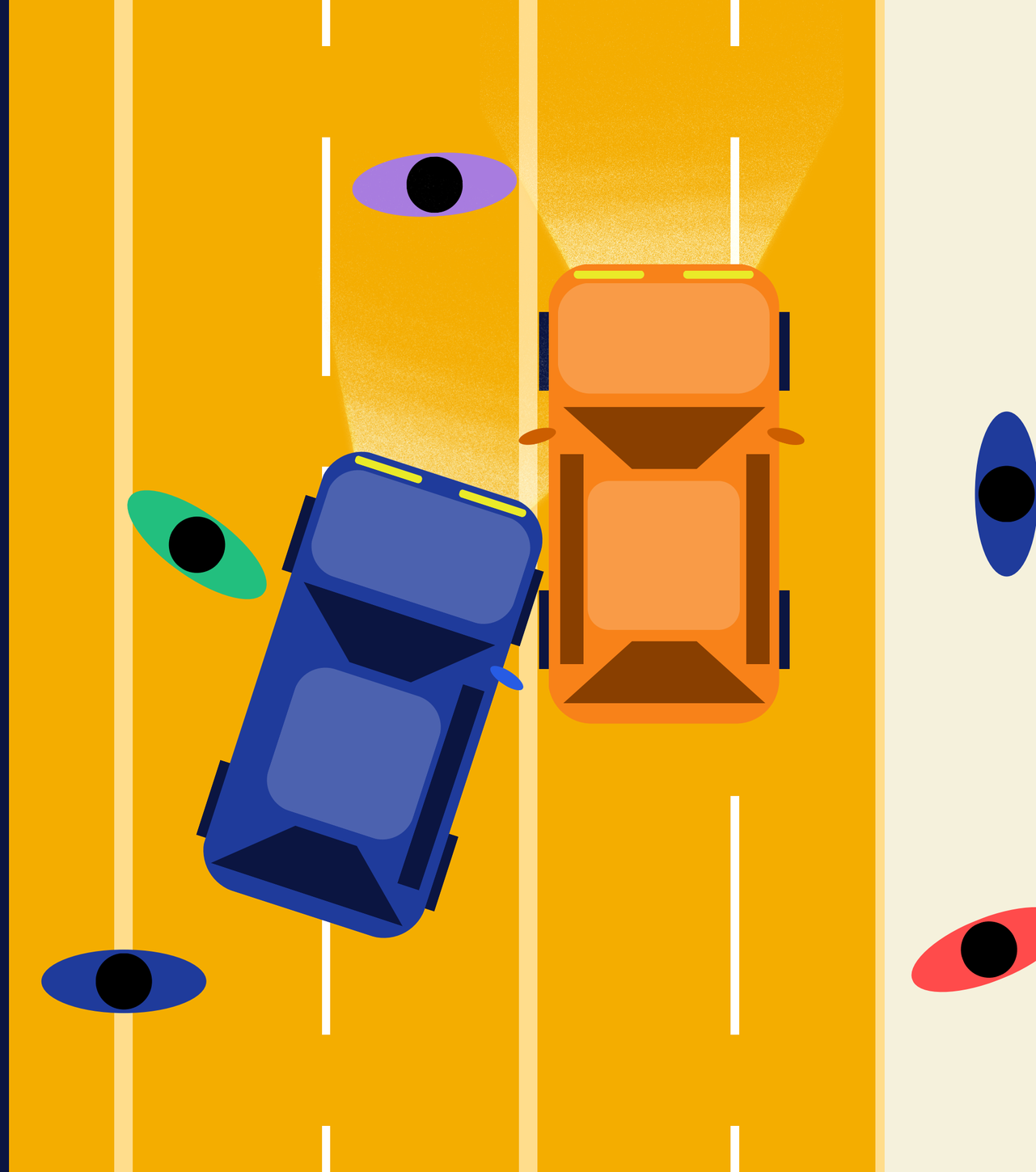
1. NGUYỄN VĨNH PHƯỚC - K214140905
2. NGUYỄN HOÀI NAM - K204020091
3. NGUYỄN QUỐC TRỌNG NGHĨA - K214140947
4. MAI MINH NHẬT - K214141336

CONTENT

DATASET INTRODUCTION

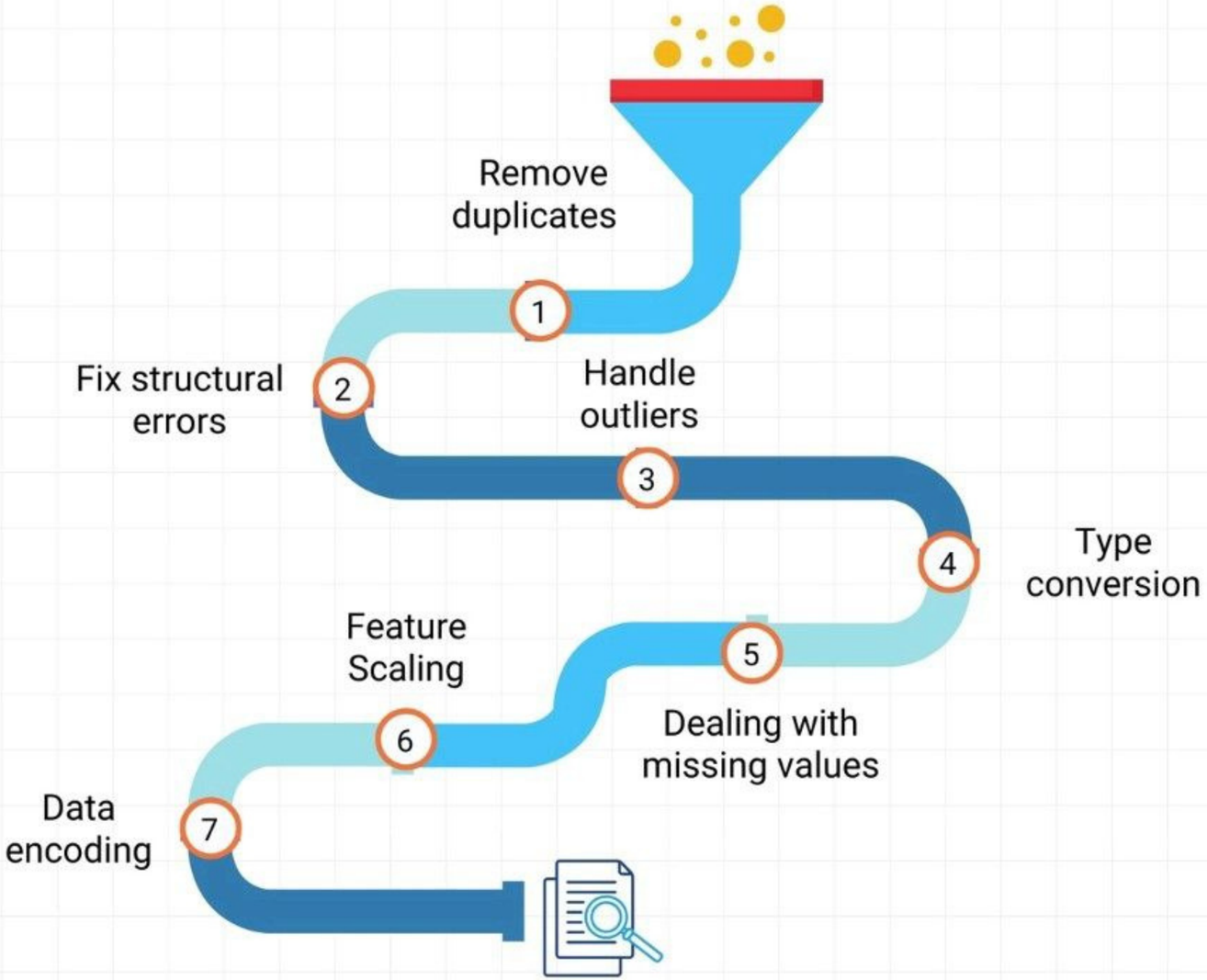
WHY CHOOSE IT

Increasingly, insurance companies are paying attention to:
Predicting customers with high risk tendencies => high probability of using insurance.
=> Maximizing profits.
=> Providing appropriate strategies.



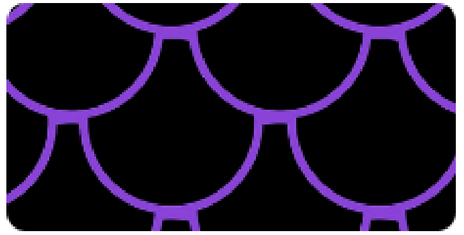
Data preprocessing

The foundation of data science solution



INTRODUCTION

Car Insurance Claim Data



Data Card | Code (5) | Discussion (3)

About Dataset

No description available

Insurance

car_insurance_claim.csv (1.58 MB)

Detail | Compact | Column | 10 of 27 columns

ID	KIDSDRIV	BIRTH	AGE	HOMEKIDS	YOJ

Usability ⓘ
2.94

License
Unknown

Expected update frequency
Not specified

Data Explorer
Version 1 (1.58 MB)

car_insurance_claim.csv

KAGGLE

INTRODUCTION

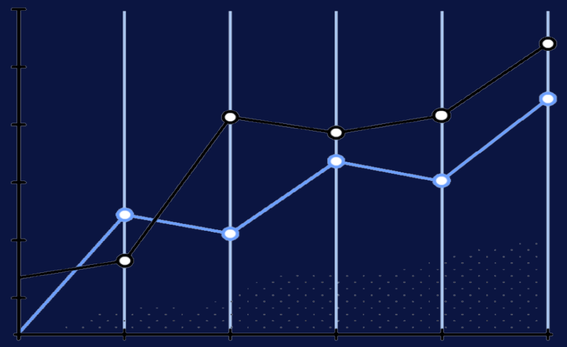
ID	CUSTOMER ID
KIDSDRIV	DRIVING TEENAGER => THE NUMBER OF TEENAGERS USING CAR
AGE	CUSTOMER AGE
HOMEKIDS	THE NUMBER OF KIDS
YOJ	YEAR ON JOB - THE LONGEST TIME CUSTOMER STAY A JOB
INCOME	SALARY EACH YEAR
PARENT1	SINGLE PARENTS (YES OR NO)
HOME_VAL	VALUE OF CUSTOMER HOUSE => IF VALUE EQUAL 0 MEANS CUSTOMER DON'T OWN ANY HOUSE
MSTATUS	MARITAL STATUS
GENDER	GENDER
EDUCATION	THE HIGHEST LEVEL OF EDUCATION
OCCUPATION	NAME OF JOB
TRAVTIME	TIME TO WORK
CAR_USE	THIS IS THE PRIVATE OR COMMERCIAL VEHICLE USED BY CUSTOMER
BLUEBOOK	THE VALUE OF CARS
TIF	TIME IN FORCE - HOW LONG THEY BECOME CUSTOMER WITH COMPANY

INTRODUCTION

CAR_TYPE	TYPE OF CAR
RED_CAR	COLOR OF CAR IS RED OR NOT
OLDCLAIM	THE PREVIOUS PAYOUTS IN LAST 5 YEARS
CLM_FREQ	TOTAL NUMBER OF CLAIMS
REVOKED	LICENSE REVOKED IN 7 YEARS
MVR_PTS	MOTOR VEHICLE RECORD POINTS => GET MORE TRAFFIC TICKETS (GIẤY XỬ PHẠT)
CLM_AMT	USED CLAIM PAYMENT
CAR_AGE	AGE OF CAR
CLAIM_FLAG	WAS CAR IN CRASH (1 OR 0)
URBANICITY	HOME/WORK AREA

10302 observations
27 variables

DATA CLEANING



CONTENT

MISSING VALUES

DATA INCONSISTENT

OUTLIERS

DATA INCONSISTENT

INCOME	EDUCATION	CAR_TYPE
$67,349	PhD	Minivan
$91,449	z_High School	Minivan
$52,881	Bachelors	Van
$16,039	z_High School	z_SUV
$114,986	<High School	Minivan
$125,301	PhD	z_SUV
$18,755		

DATA INCONSISTENT

```
df['INCOME'] = df['INCOME'].str.replace('$', '').str.replace(',', '').astype(float)
df['HOME_VAL'] = df['HOME_VAL'].str.replace('$', '').str.replace(',', '').astype(float)
df['OLDCLAIM'] = df['OLDCLAIM'].str.replace('$', '').str.replace(',', '').astype(float)
df['CLM_AMT'] = df['CLM_AMT'].str.replace('$', '').str.replace(',', '').astype(float)
df['GENDER'] = df['GENDER'].str.replace('z_', '')
df['MSTATUS'] = df['MSTATUS'].str.replace('z_', '')
df['EDUCATION'] = df['EDUCATION'].str.replace('z_','').str.replace('<','')
df['OCCUPATION'] = df['OCCUPATION'].str.replace('z_','')
df['BLUEBOOK'] = df['BLUEBOOK'].str.replace('$','').str.replace(',','').astype(int)
df['CAR_TYPE'] = df['CAR_TYPE'].str.replace('z_','')
df['CAR_TYPE'] = df['CAR_TYPE'].str.replace('z_','')
df['URBANICITY'] = df['URBANICITY'].str.replace('Highly Urban/ ','').str.replace('z_Highly Rural/ ','')
```

MISSING VALUES

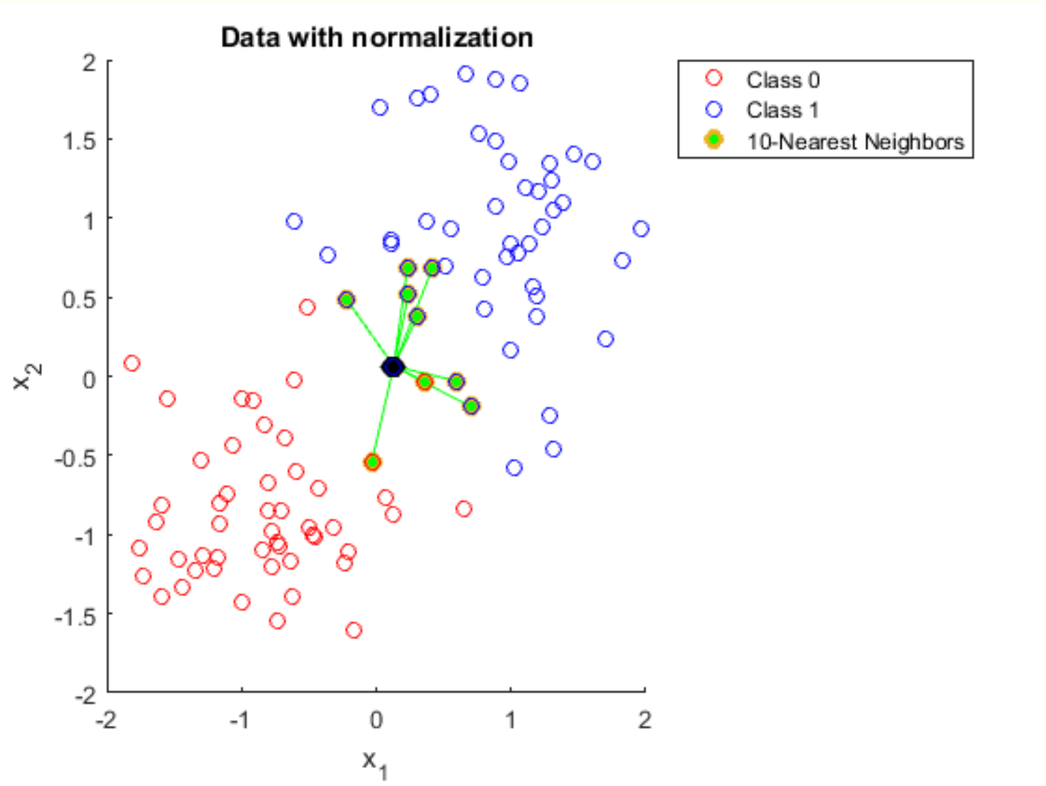
MODE IMPUTATION

Make
Ford
Ford
Fiat
BMW
Ford
Kia
Fiat
Ford
Kia

Mode = Ford

Price
Ford
Ford
Fiat
BMW
Ford
Kia
Ford
Fiat
Ford
Ford
Kia

KNN IMPUTATION



MISSING VALUES

```
missing_value_table(df)
```

```
These dataframe has 26 columns
These dataframe has 6 columns which has null values
```

	Missing Value	Percentage of Missing
OCCUPATION	665	6.5
CAR_AGE	639	6.2
HOME_VAL	575	5.6
INCOME	570	5.5
YOJ	548	5.3
AGE	7	0.1

CATEGORY

OCCUPATION

NUMERIC

CAR_AGE
HOME_VAL
INCOME
YOJ
AGE

OUTLIERS

417 OUTLIERS

EXPLORATORY DATA ANALYSIS

CONTENT

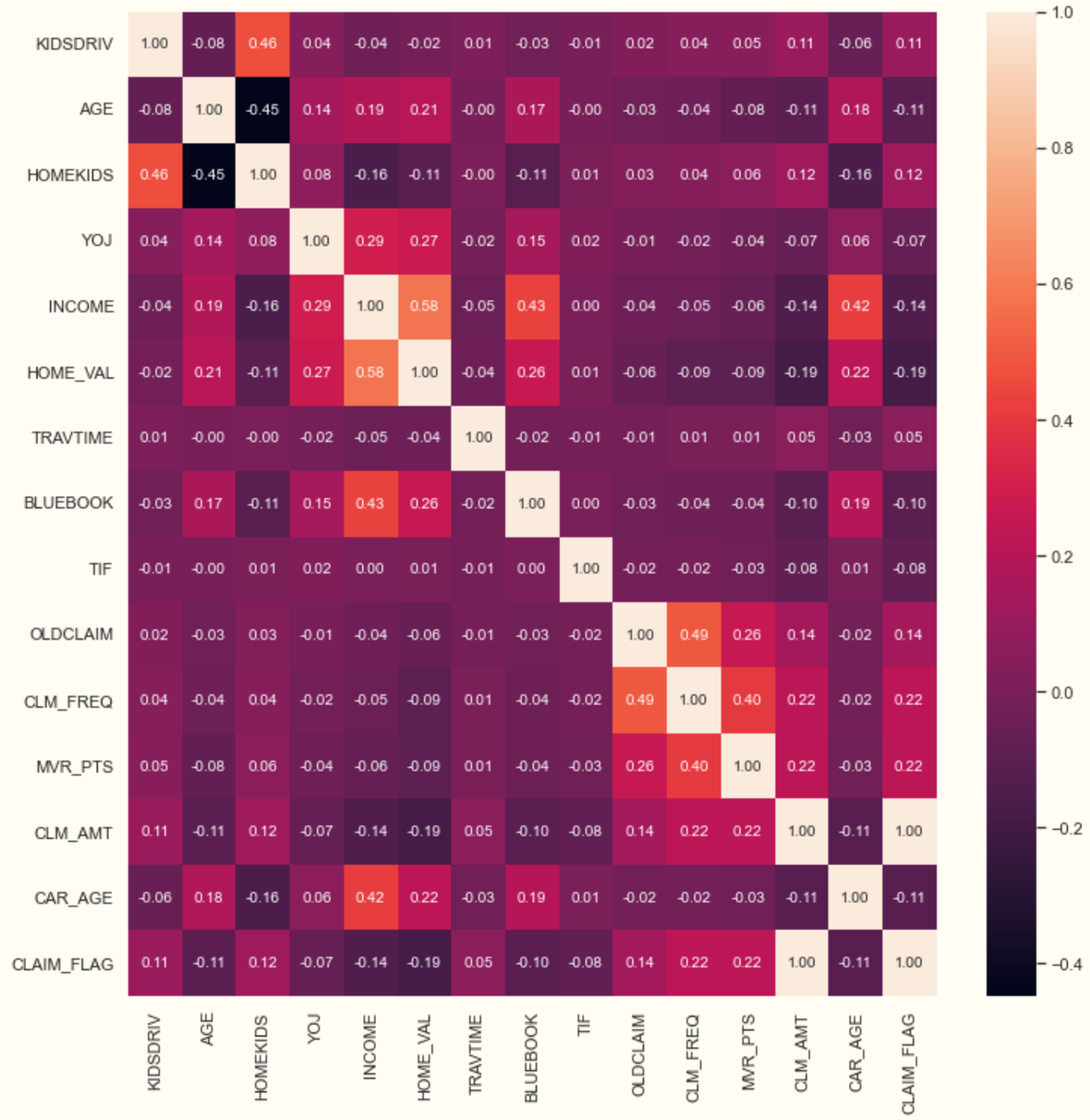
HISTORY OF USING CAR INSURANCE

CUSTOMER PROFILE

VEHICLE

CUSTOMER'S PROFILE

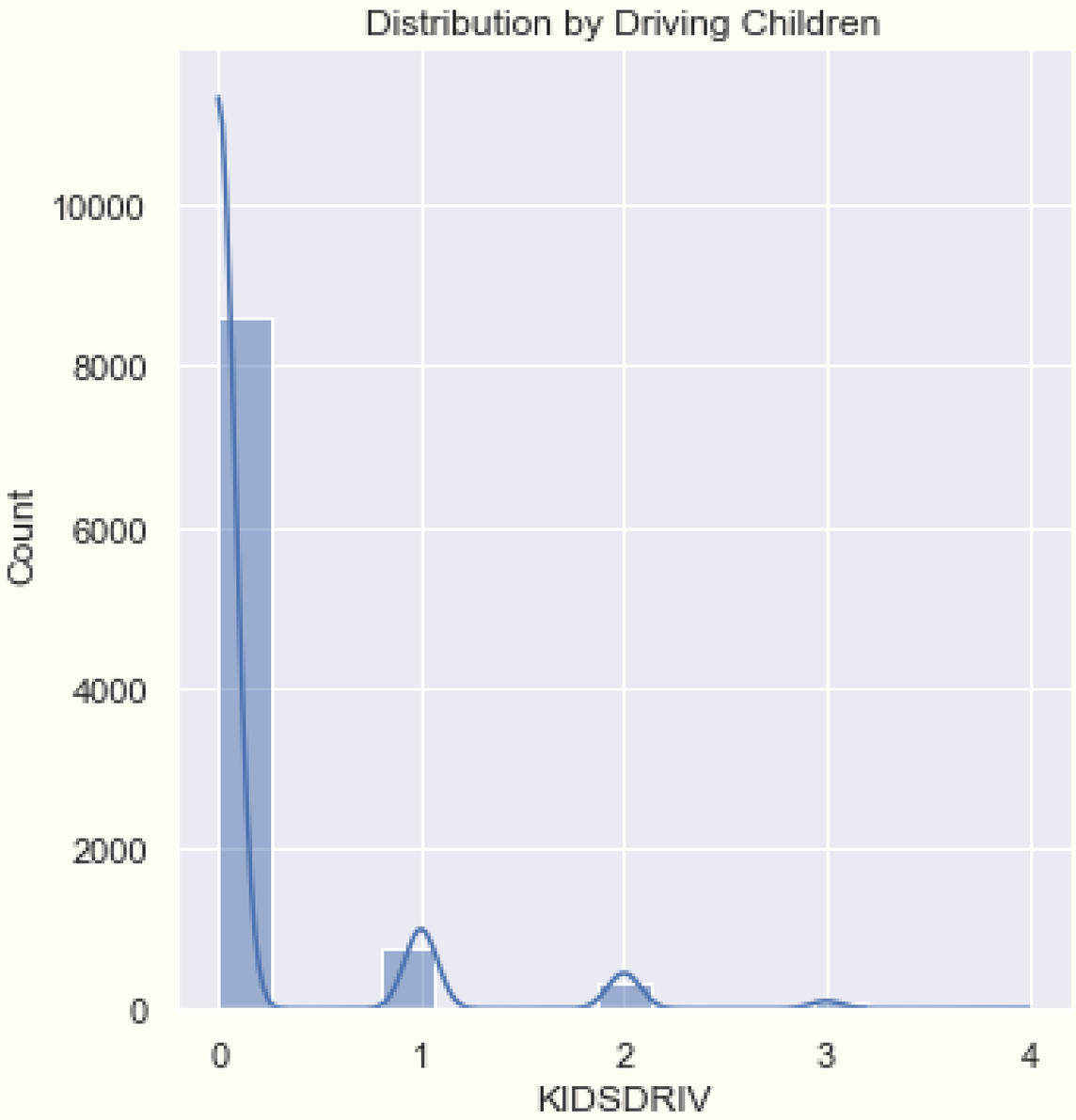
CORRELATION



The correlation between CLM_AMT and CLAIM_FLAG is 1

ALL CUSTOMERS INVOLVED IN CAR ACCIDENTS USE INSURANCE.

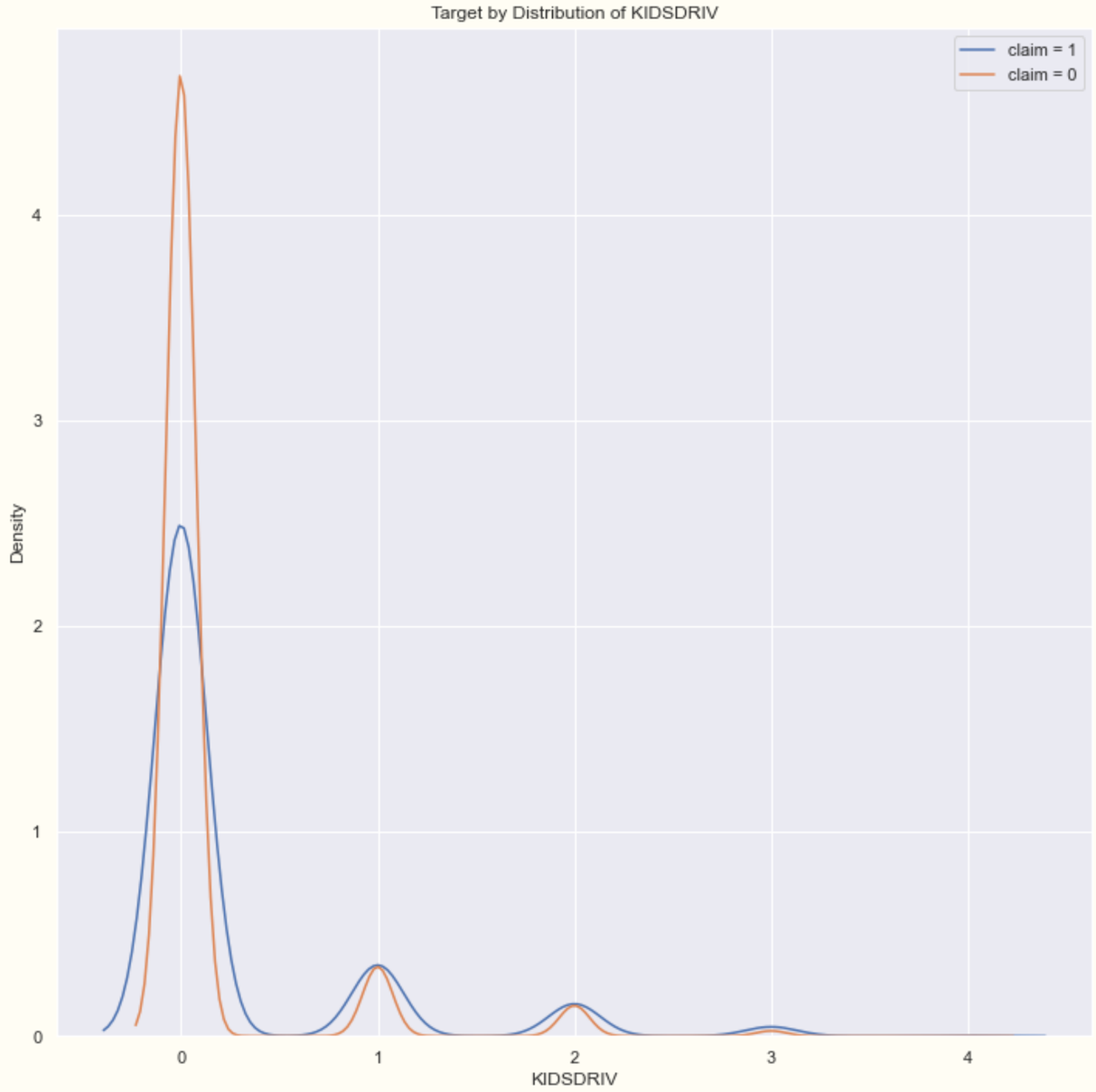
KIDS DRIVING



83.7%

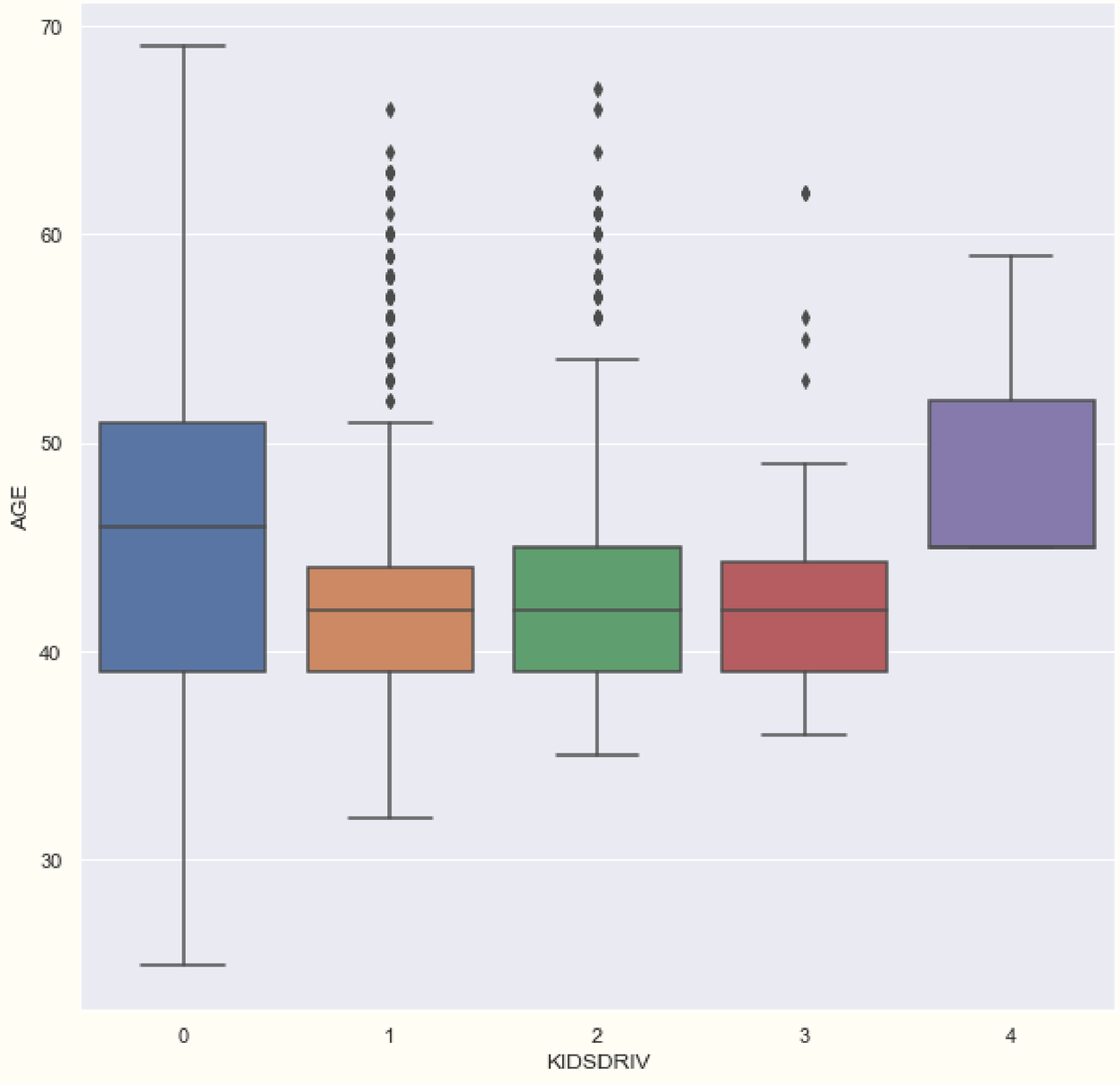
The group of customers who do not allow teenagers to use their cars have the highest proportion

KIDS DRIVING



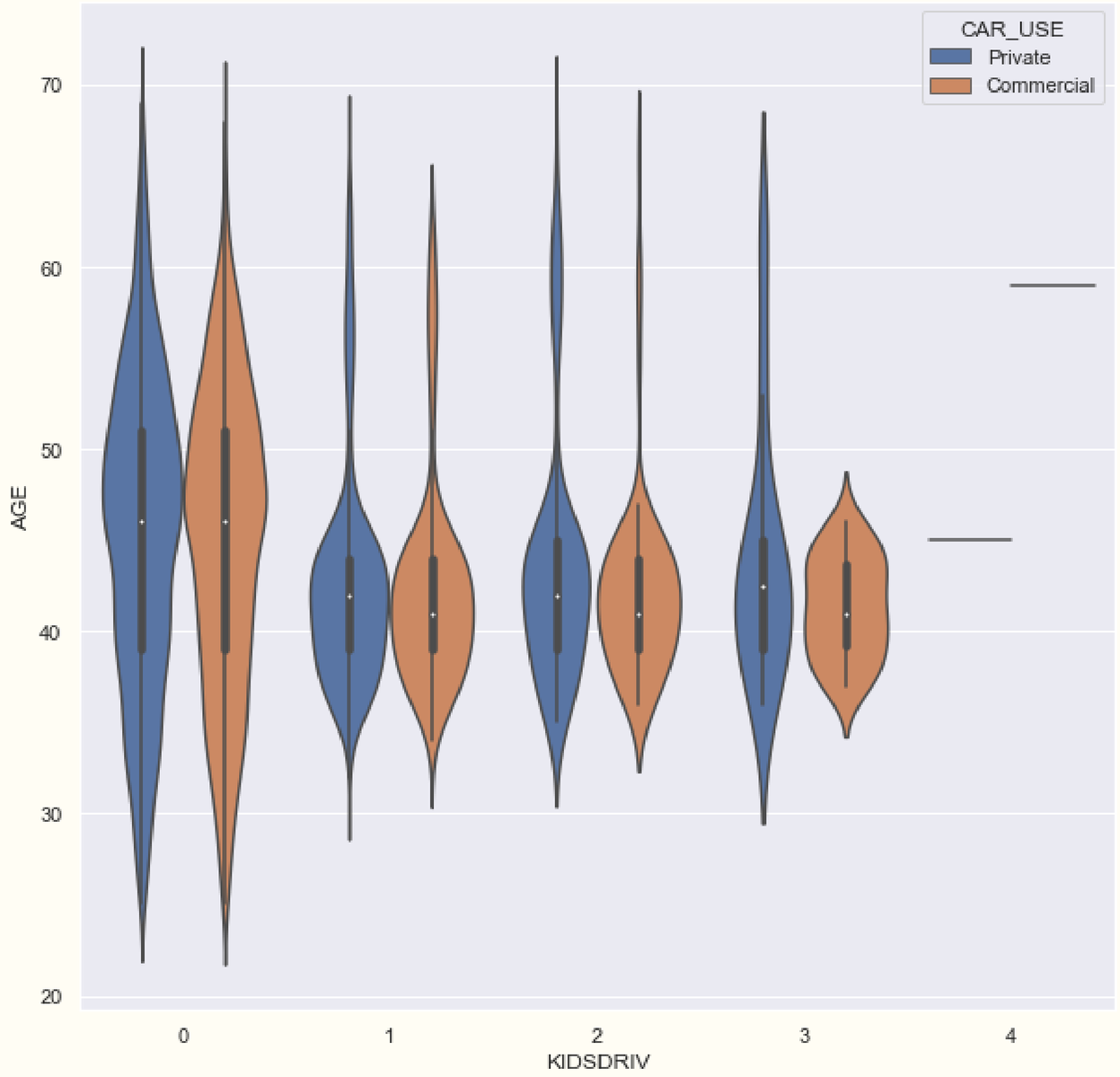
THE GROUP OF CUSTOMERS WHO ALLOW TEENAGERS TO USE THEIR VEHICLES TEND TO HAVE A HIGHER TENDENCY TO USE CAR INSURANCE.

KIDS DRIVING



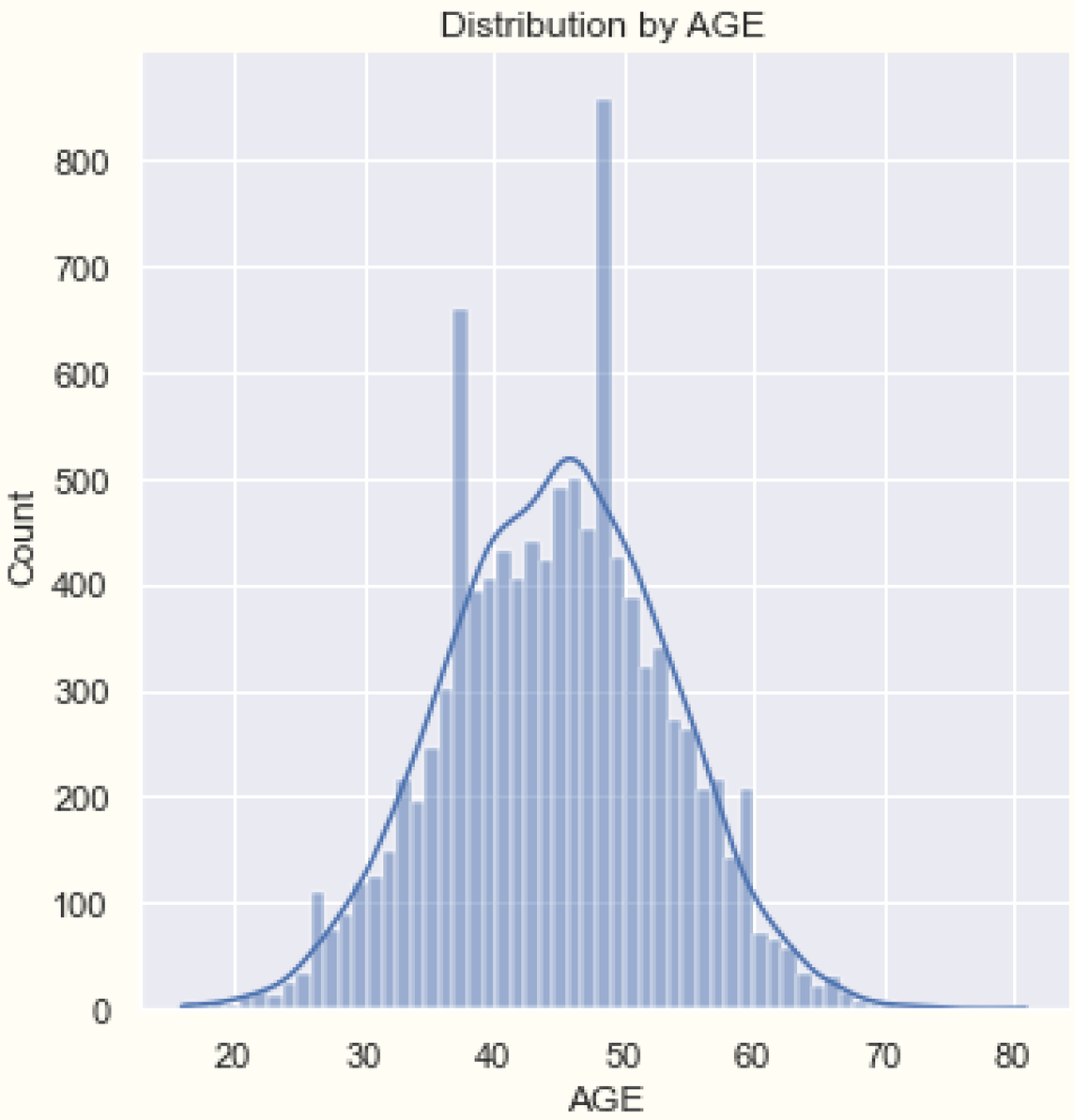
THE AGE GROUP OF CUSTOMERS ALLOWING TEENAGERS TO USE THEIR VEHICLES IS CONCENTRATED FROM 35-55.

KIDS DRIVING



THERE IS NO AGE DIFFERENCE BETWEEN PERSONAL OR COMMERCIAL VEHICLES WITHIN THE KIDSDRIV GROUP

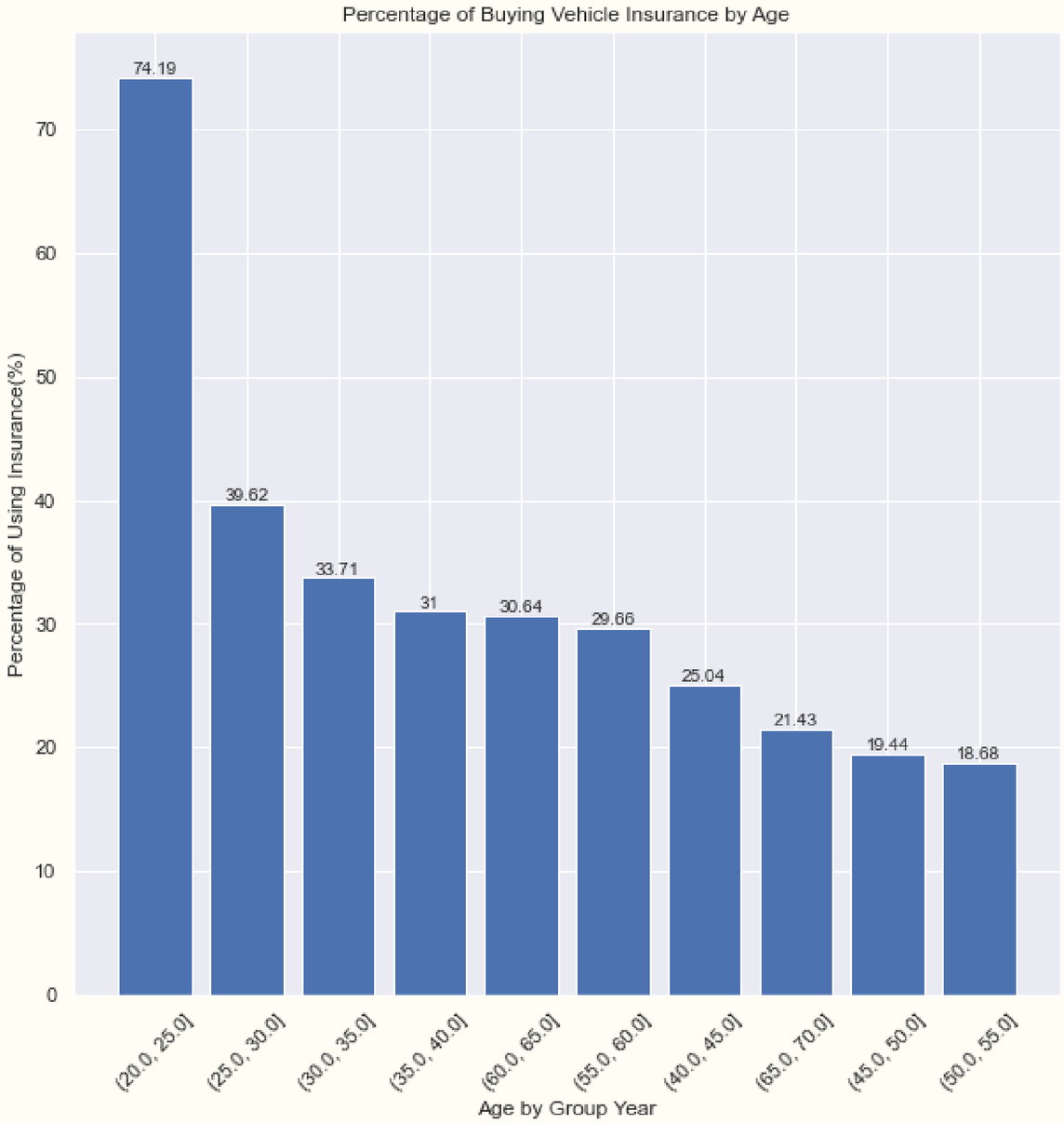
AGE



35-55

THE MAIN AGE GROUP OF THE DATA

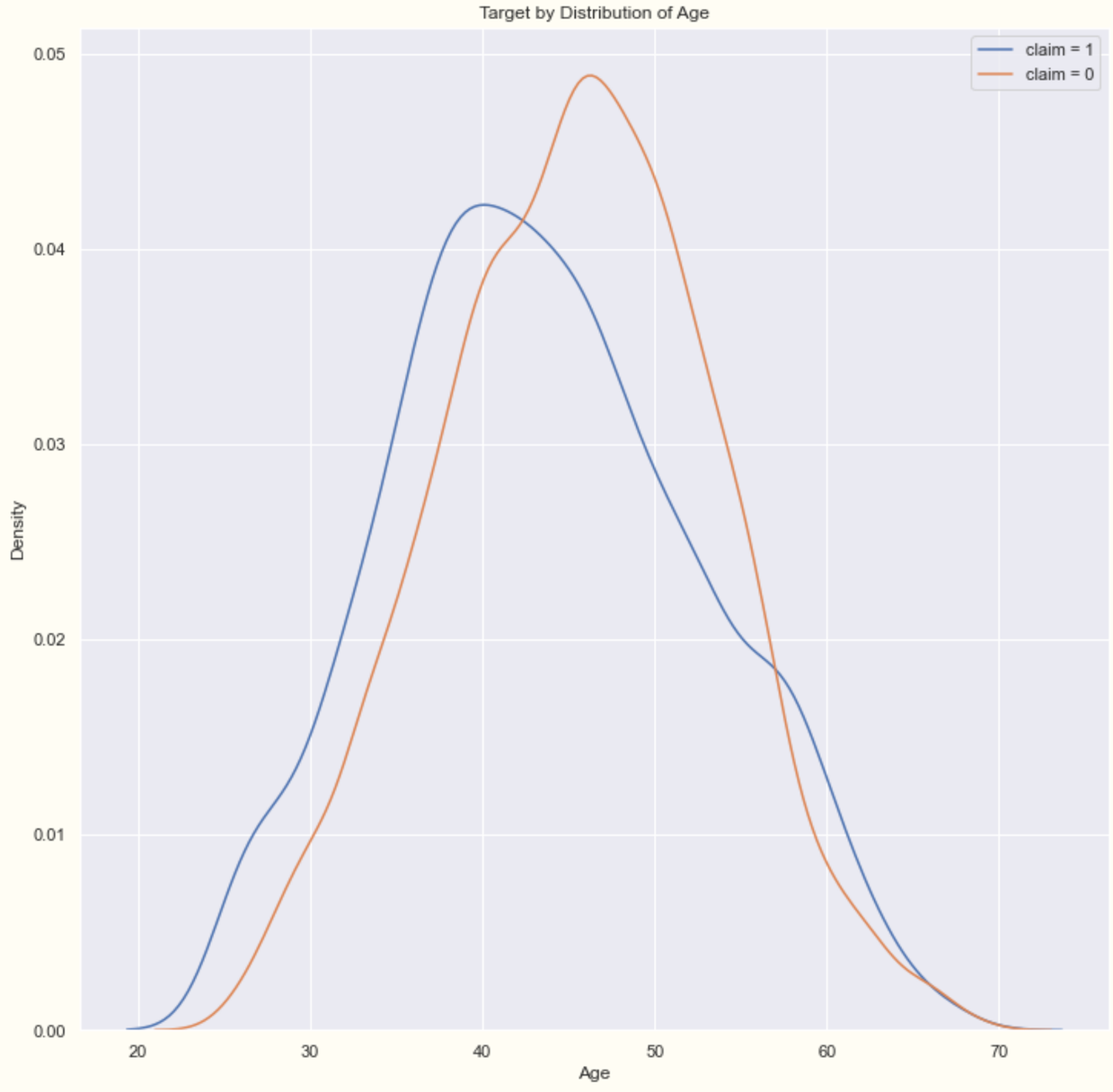
AGE



19%-31%

PERCENTAGE OF THE AGE GROUP 35-55 USING CAR INSURANCE

AGE

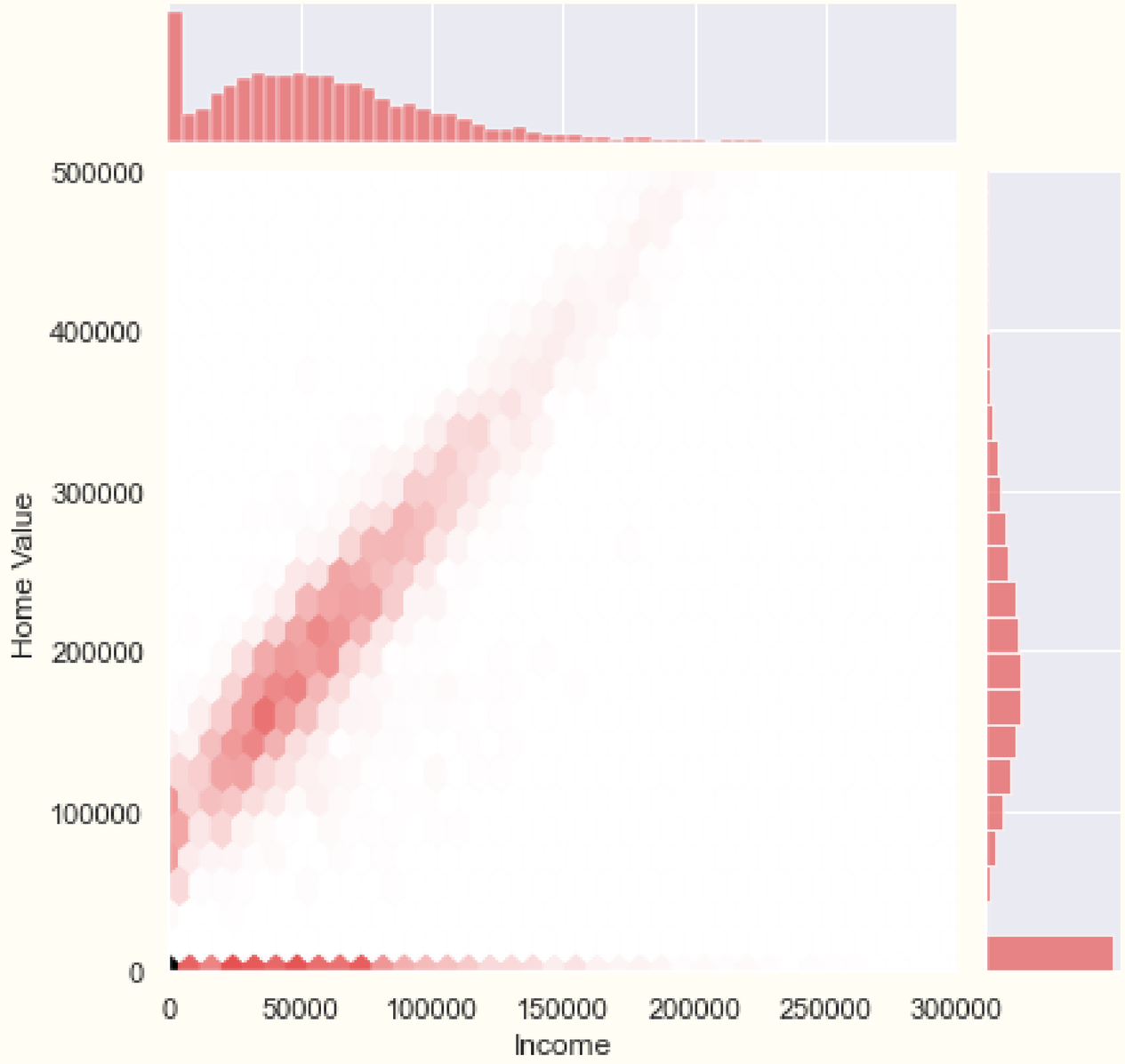


THE AGE GROUPS THAT MAKE UP A LARGE NUMBER OF THE DATASET ARE ALSO THE GROUPS WITH THE HIGHEST CAR INSURANCE USAGE

INCOME AND HOME VALUE

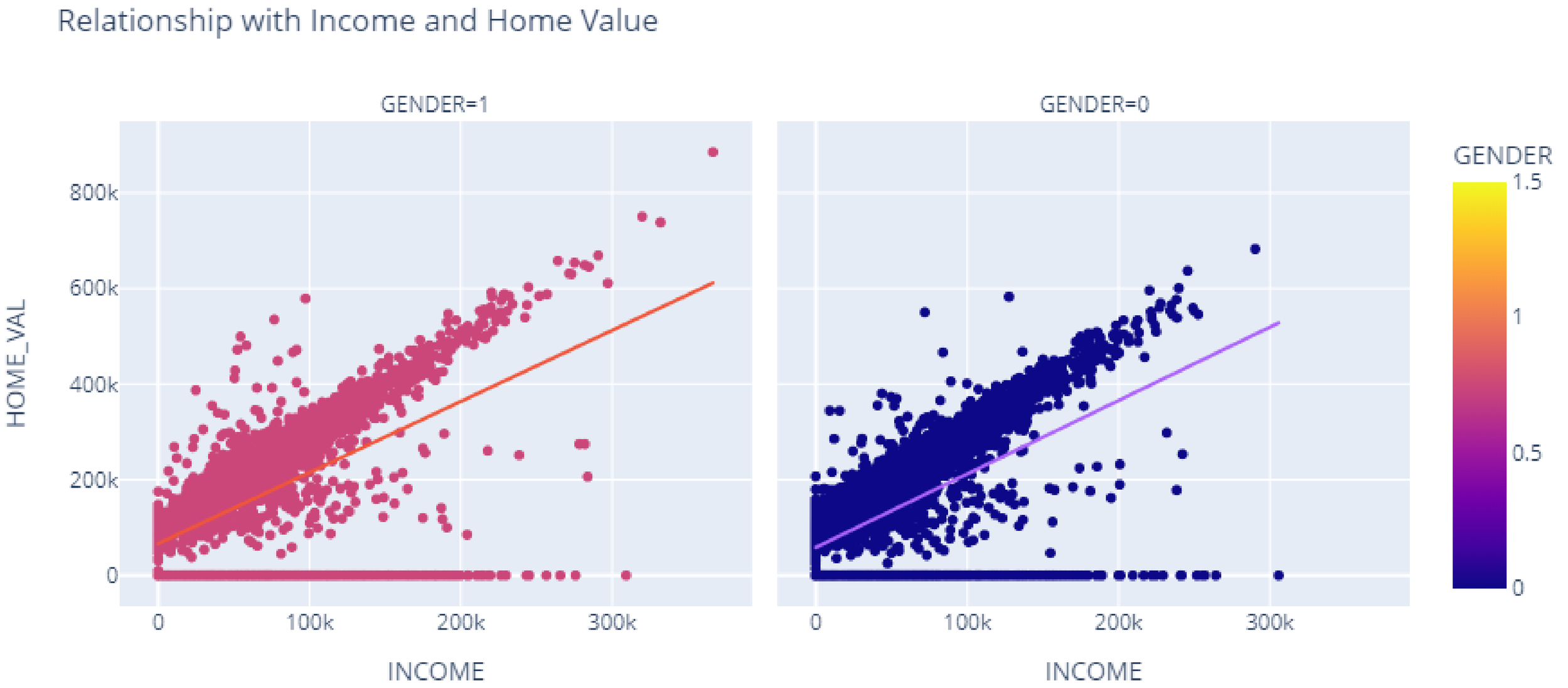
THE RELATIONSHIP BETWEEN INCOME AND HOME VALUE IS POSITIVE.

INCOME AND HOME VALUE

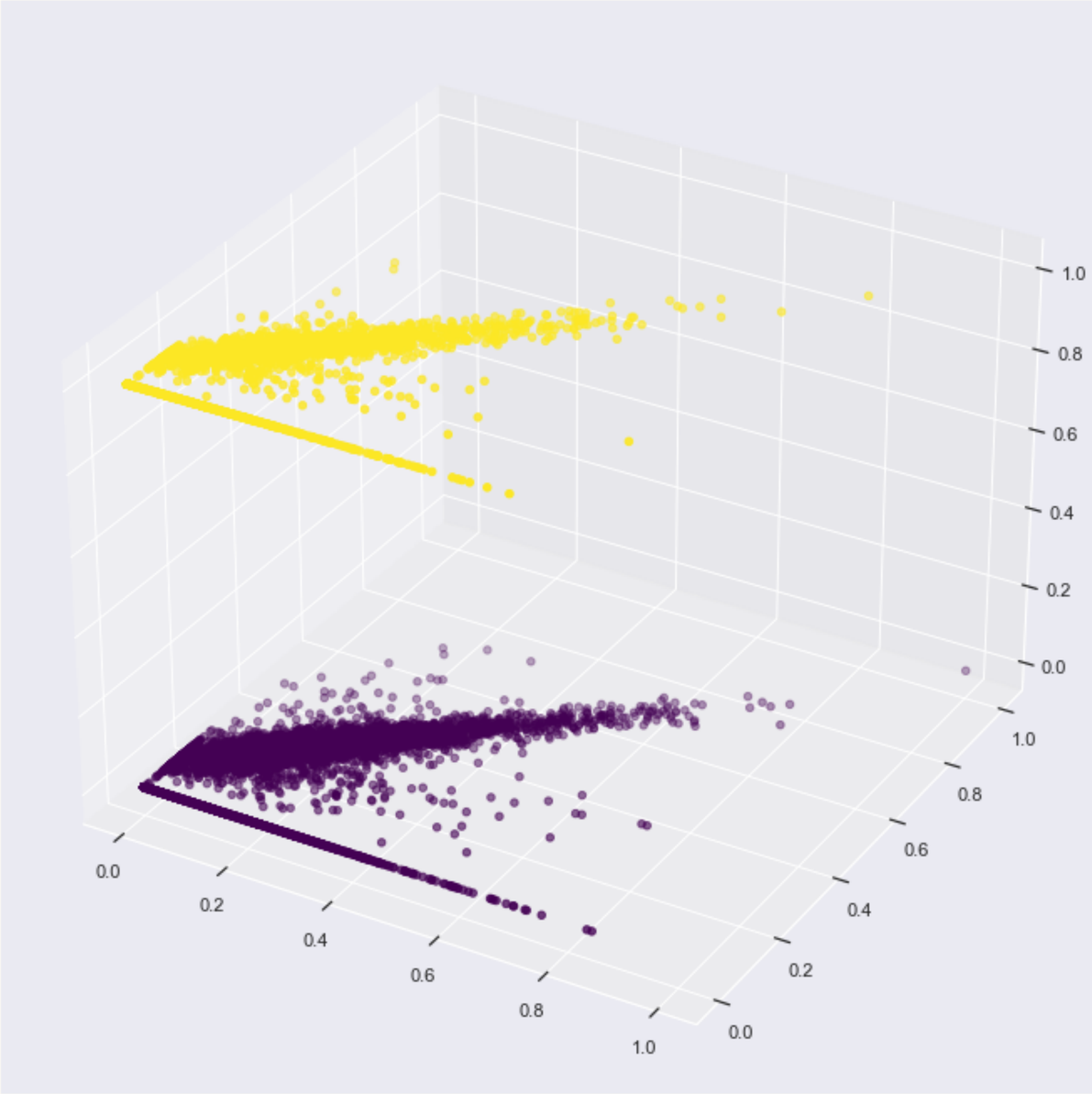


THE RELATIONSHIP BETWEEN INCOME AND HOME VALUE IS POSITIVE.

INCOME AND HOME VALUE



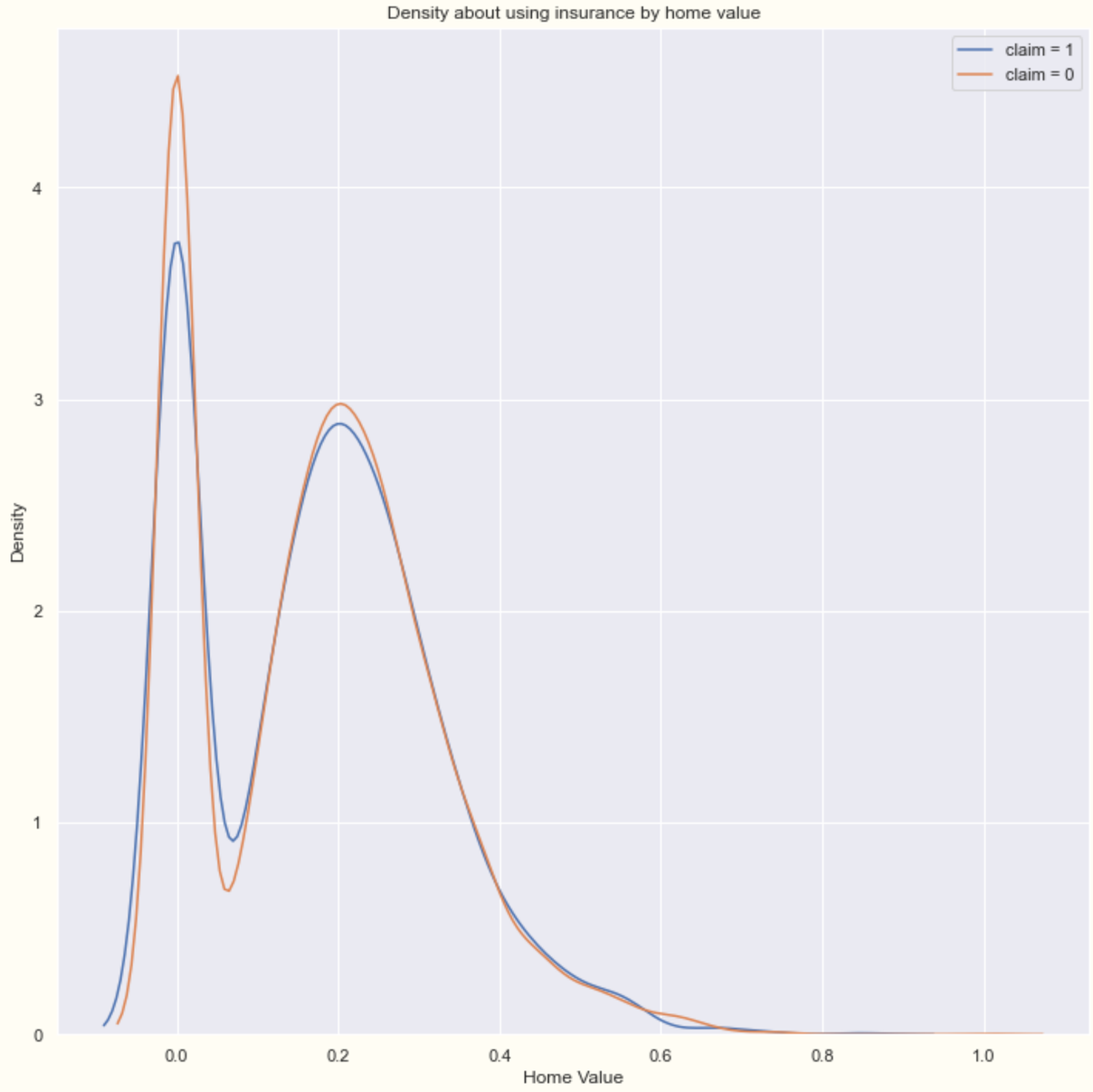
INCOME AND HOME VALUE



THERE IS VIRTUALLY NO DIFFERENCE IN THE IMPACT OF INCOME AND HOME VALUE ON THE USE OF INSURANCE

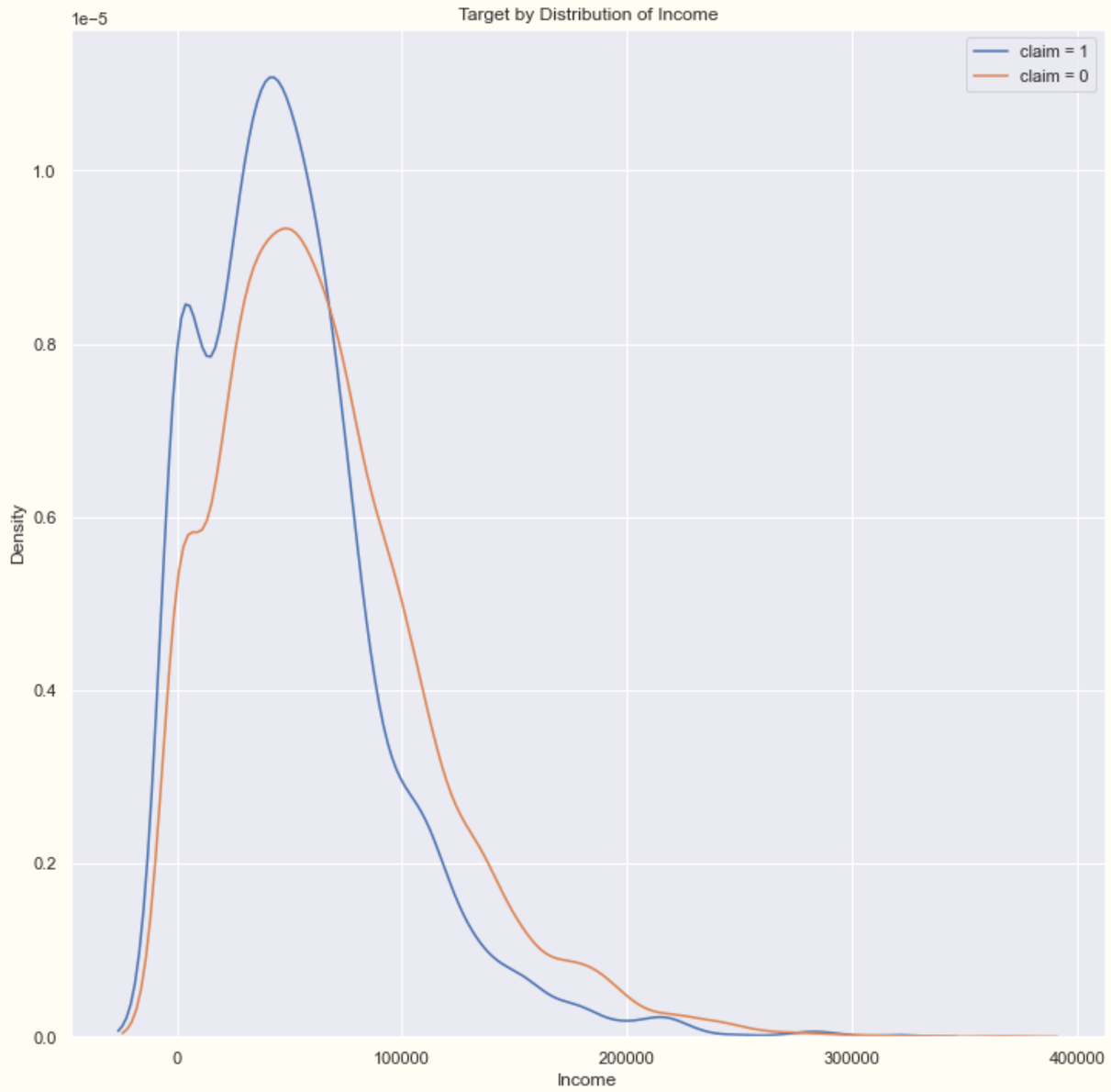
THE DATA POINTS SHOW THAT INDIVIDUALS WITH HIGH INCOME AND HOME VALUES OFTEN DO NOT USE INSURANCE.

INCOME AND HOME VALUE



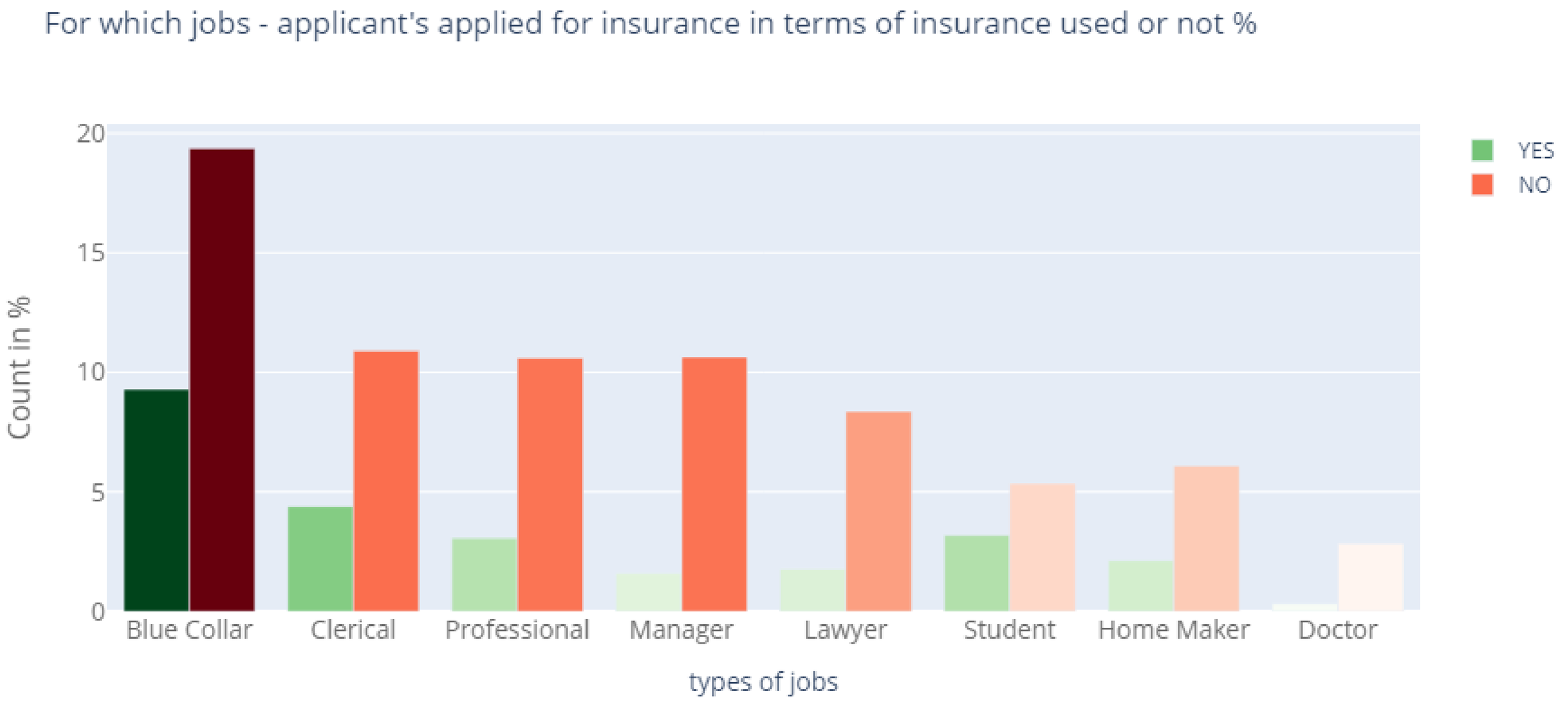
THE GROUP OF CUSTOMERS WHO DO NOT OWN HOMES TEND TO HAVE A HIGHER TENDENCY TO USE INSURANCE THAN THE GROUP OF CUSTOMERS WHO DO OWN HOMES.

INCOME AND HOME VALUE

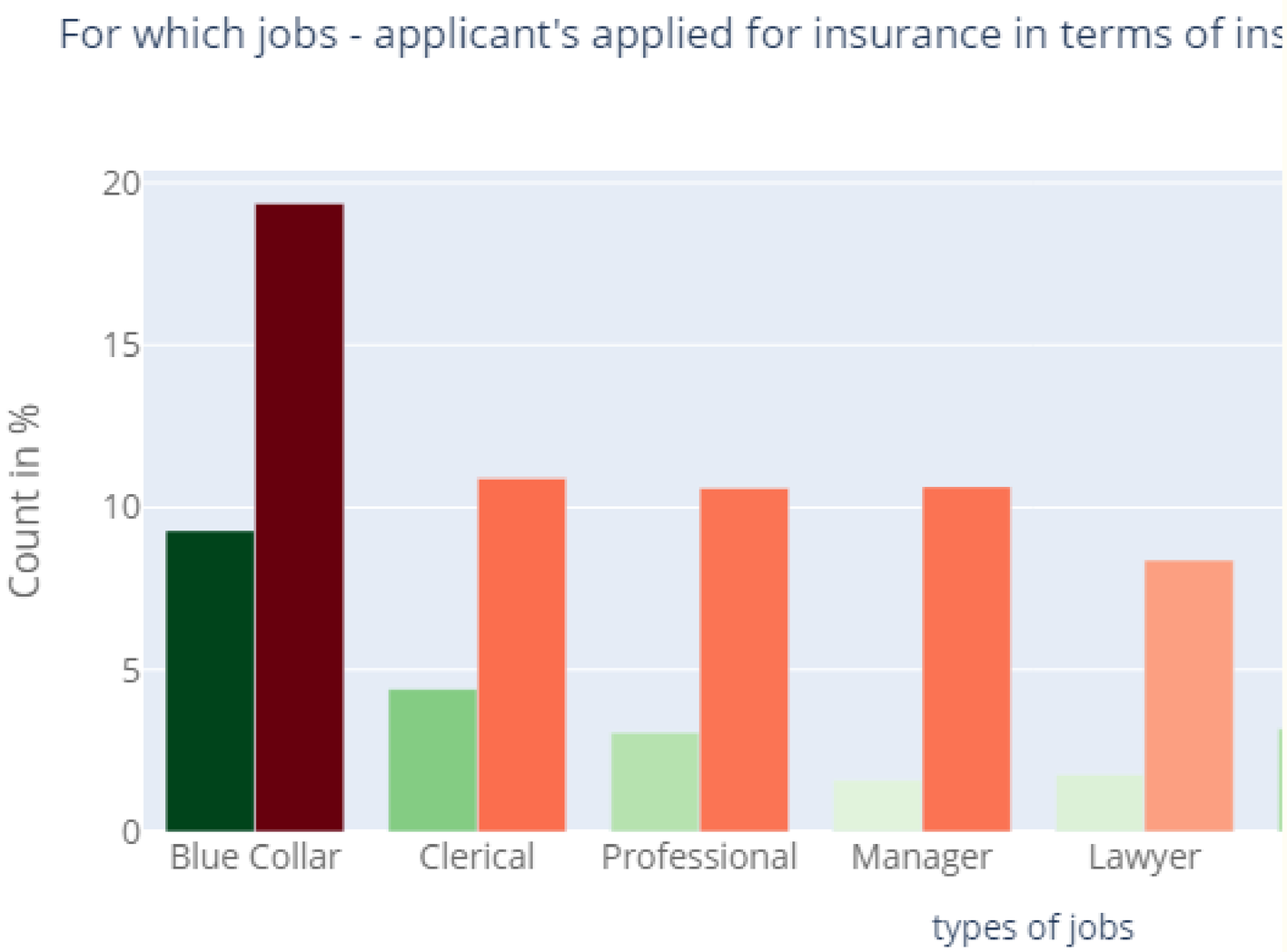


THE GROUP OF CUSTOMERS WITH LOW INCOME TEND TO HAVE A HIGHER TENDENCY TO USE INSURANCE BUT IT'S NOT CLEAR

OCCUPATION


For which jobs - applicant's applied for insurance in terms of insurance used or not %
Count in %
20
15
10
5
0
Blue Collar
Clerical
Professional
Manager
Lawyer
Student
Home Maker
Doctor
types of jobs
YES
NO

OCCUPATION



9.3%

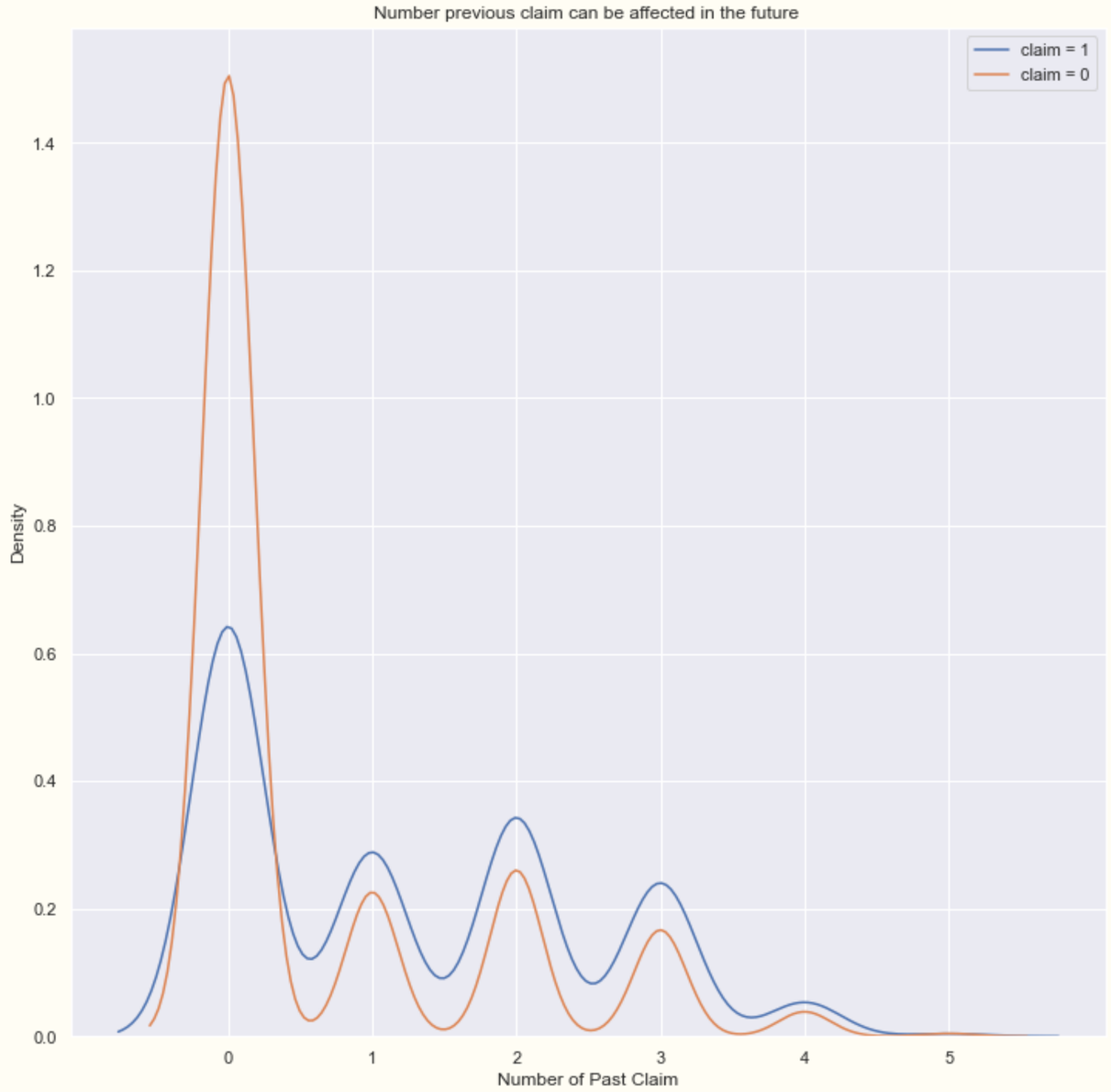
THE CUSTOMER GROUP WITH FEWER OFFICE JOBS HAS THE HIGHEST INSURANCE USAGE RATE

48.0%

CUSTOMER'S HISTORY

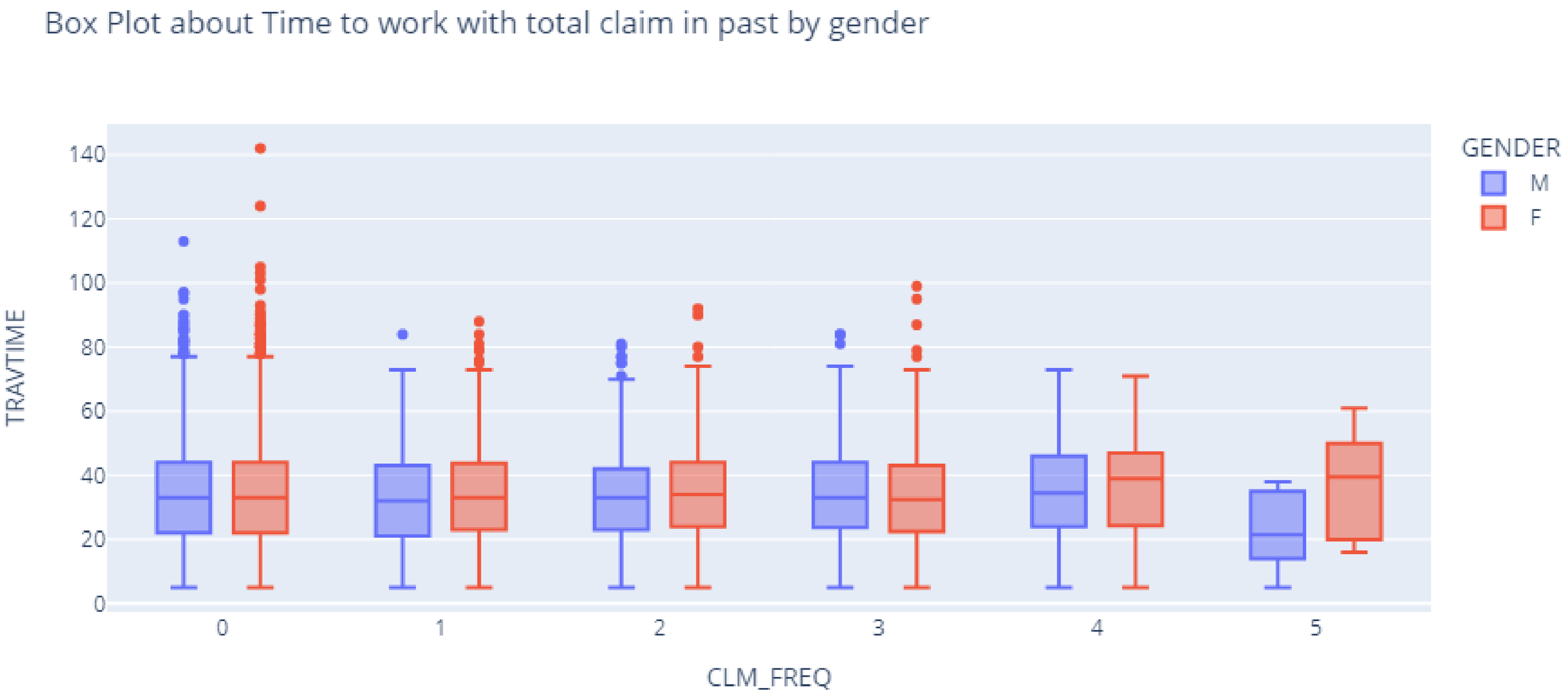
HISTORY

PREVIOUS CLAIM FREQUENCE



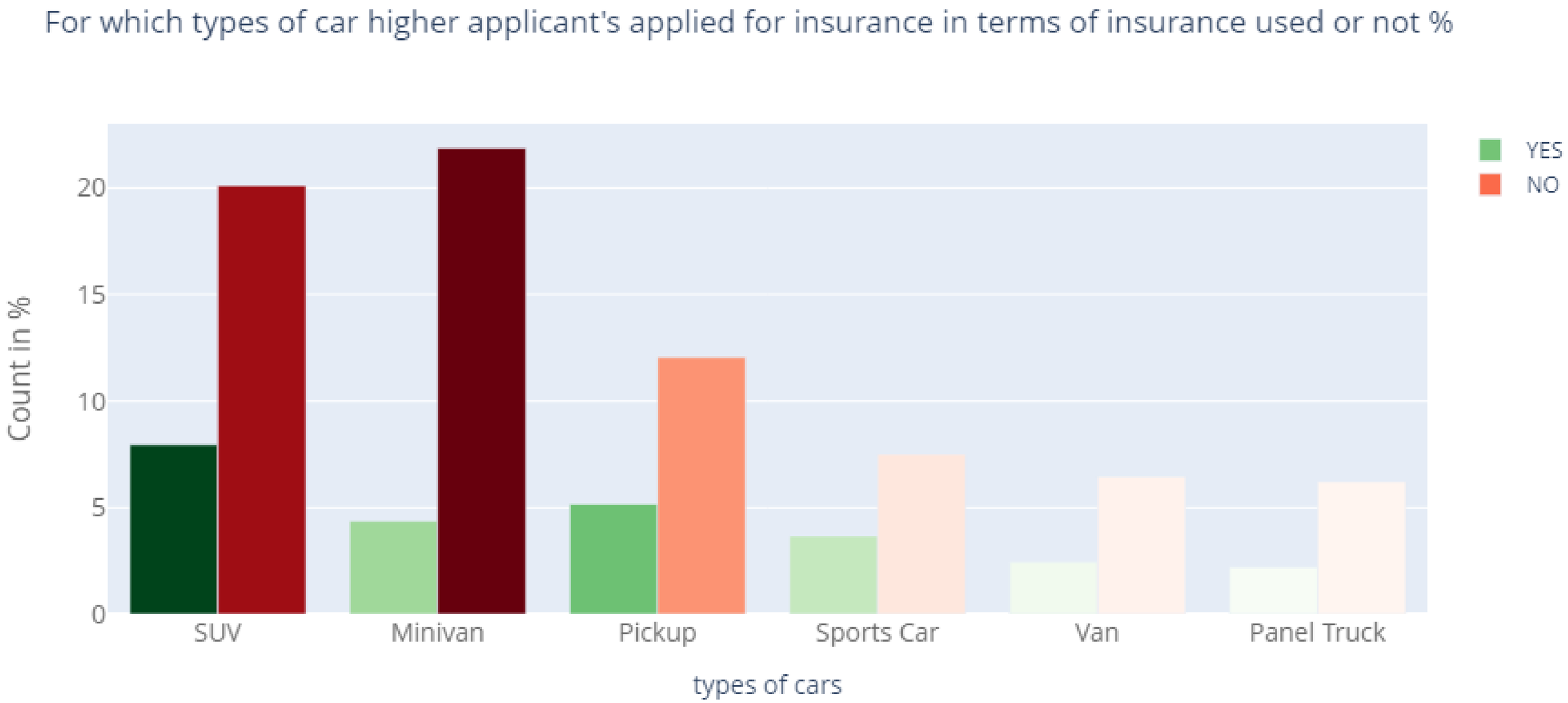
THE GROUP OF CUSTOMERS WHO HAVE A HISTORY OF USING INSURANCE HAVE A MUCH HIGHER LIKELIHOOD OF CONTINUING TO USE IT IN THE FUTURE COMPARED TO THOSE WHO HAVE NEVER USED IT BEFORE.

PREVIOUS CLAIM FREQUENCE

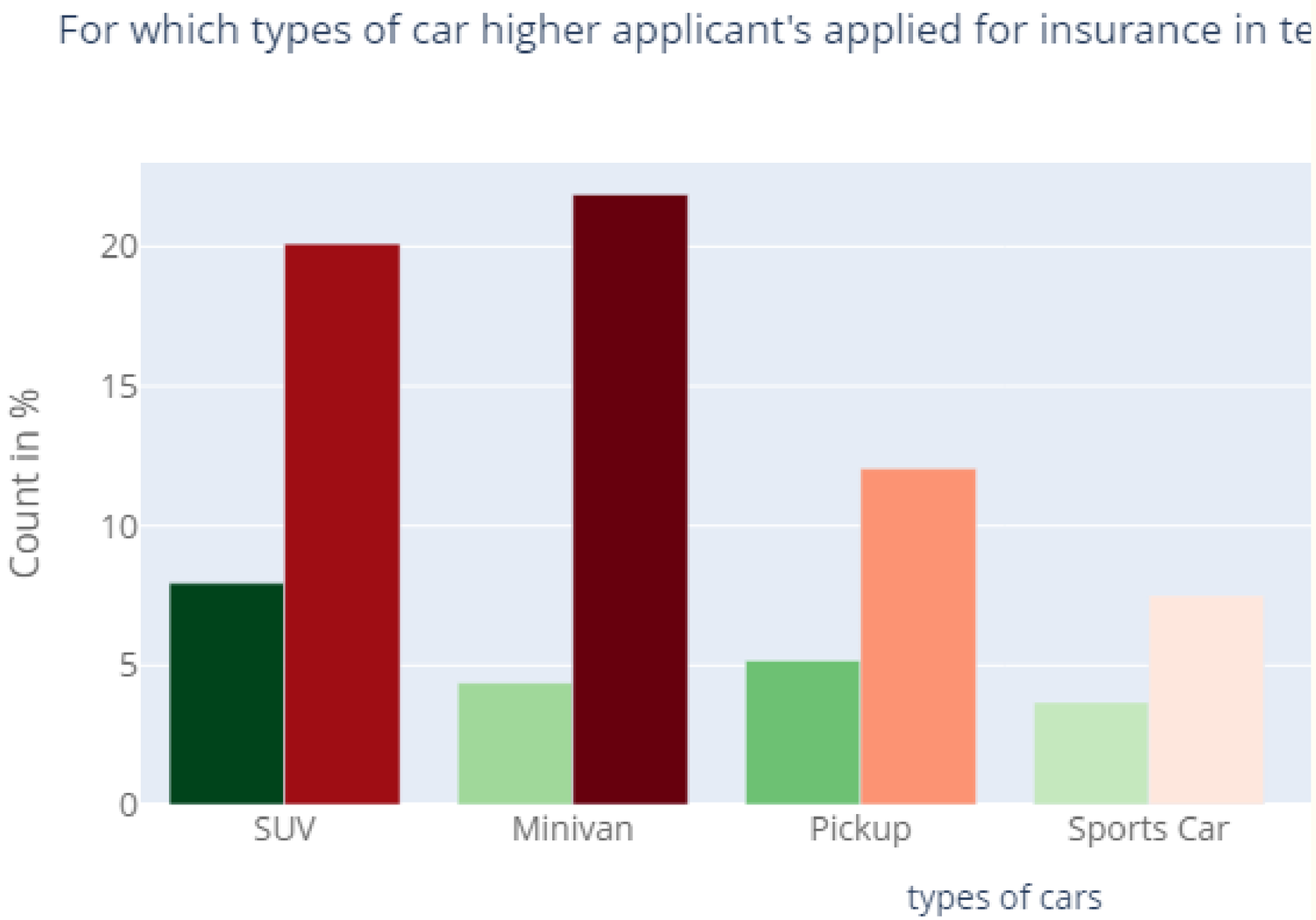


VEHICLE

TYPE OF CARS


For which types of car higher applicant's applied for insurance in terms of insurance used or not %
YES
NO
Count in %
0
5
10
15
20
SUV
Minivan
Pickup
Sports Car
Van
Panel Truck
types of cars

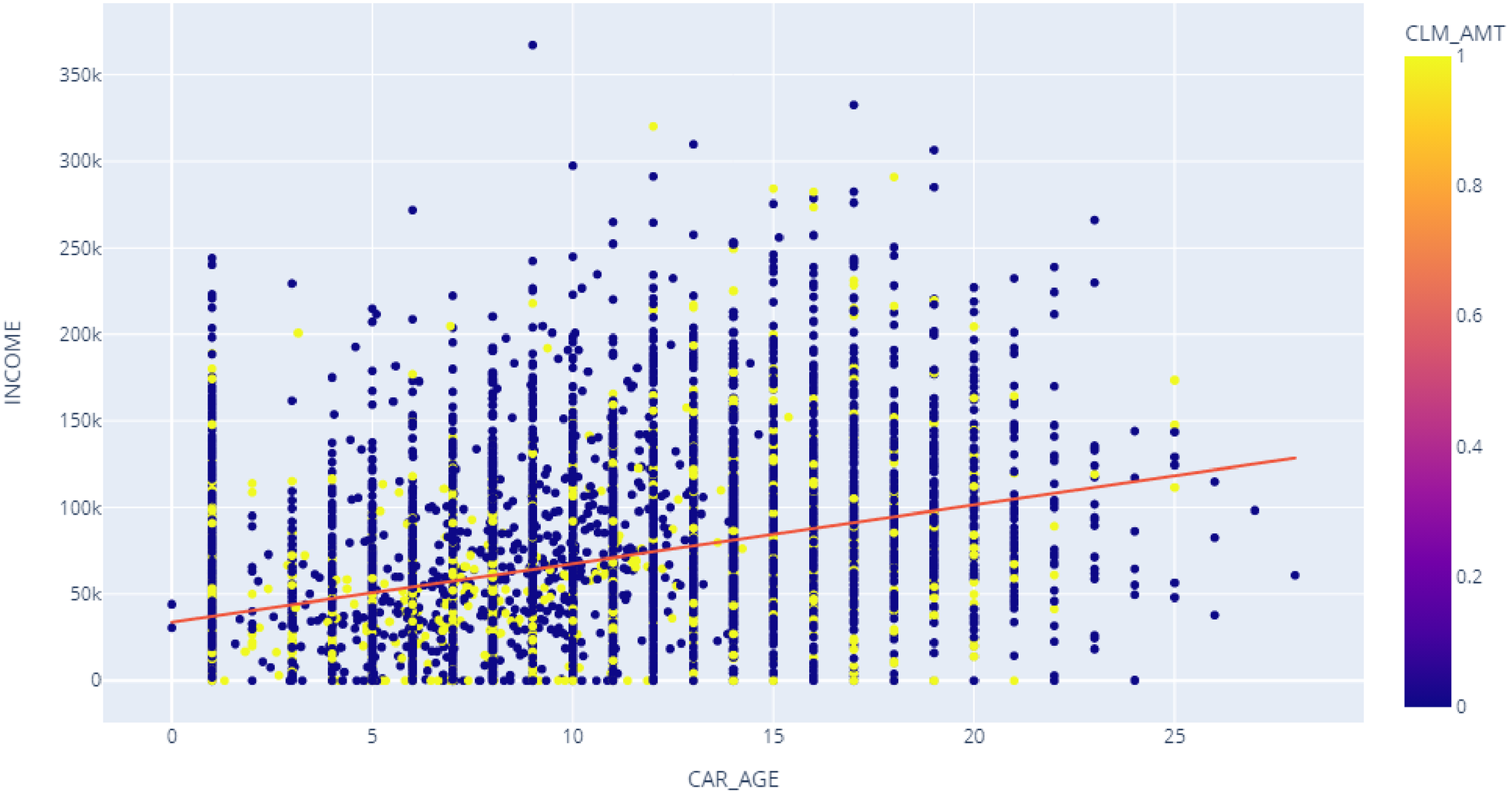
TYPE OF CARS



7.9%

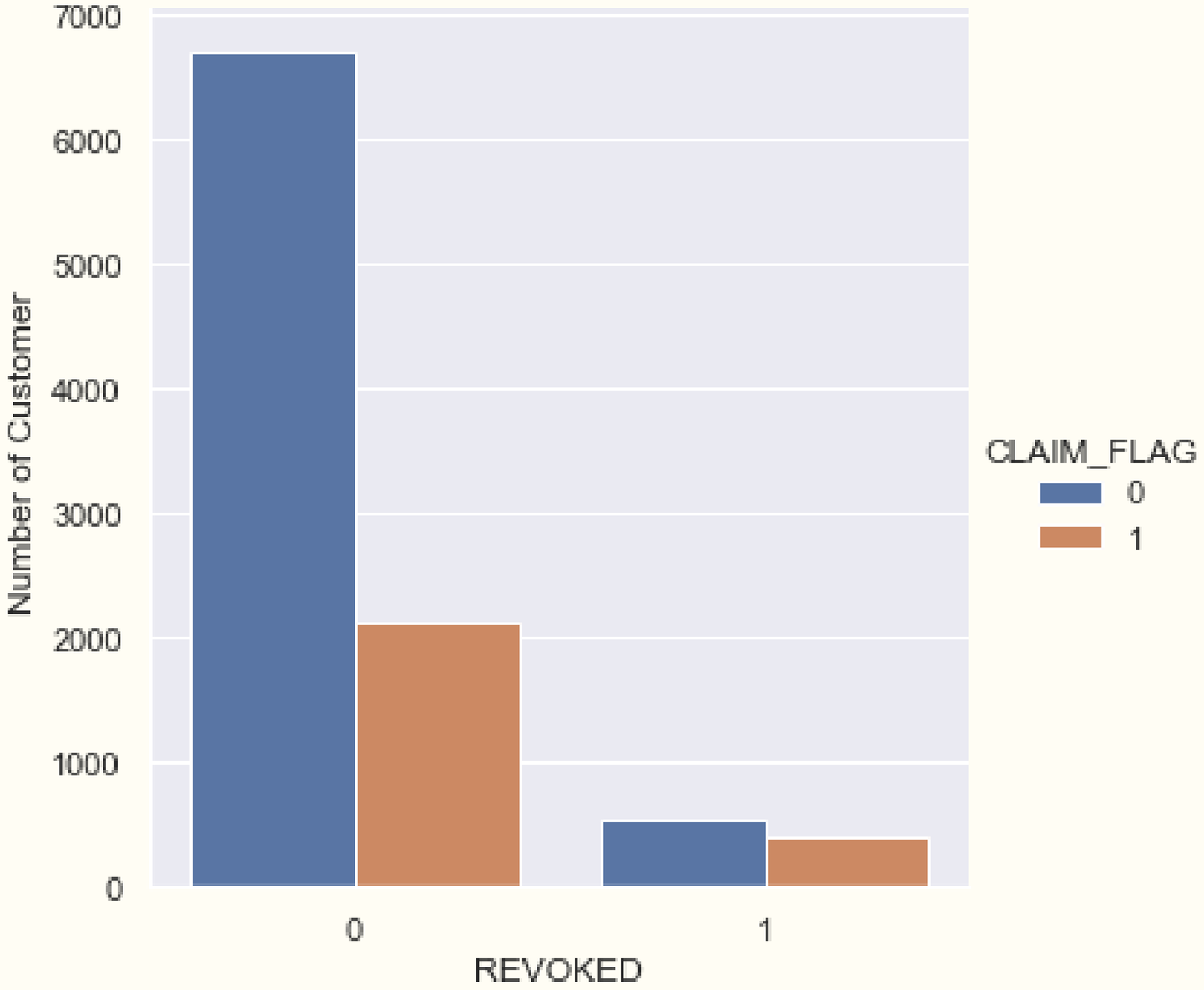
THE GROUP OF CUSTOMERS USING SUV CARS HAS THE HIGHEST BUYING RATE

CAR AGE



THE OLDER THE CAR, THE LOWER THE INSURANCE USAGE RATE

REVOKED

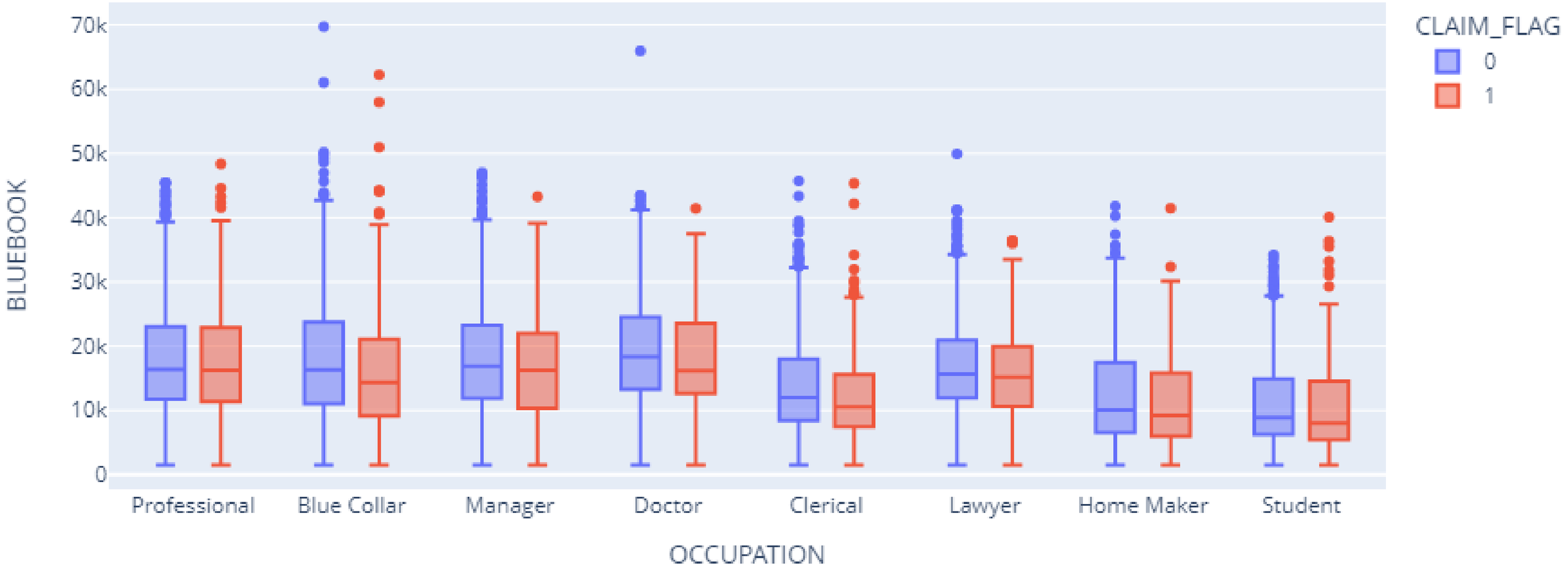


50-50

THE GROUP OF CUSTOMERS WHO HAVE HAD THEIR LICENSE REVOKED IN THE PAST TEND TO USE INSURANCE IN THE FUTURE

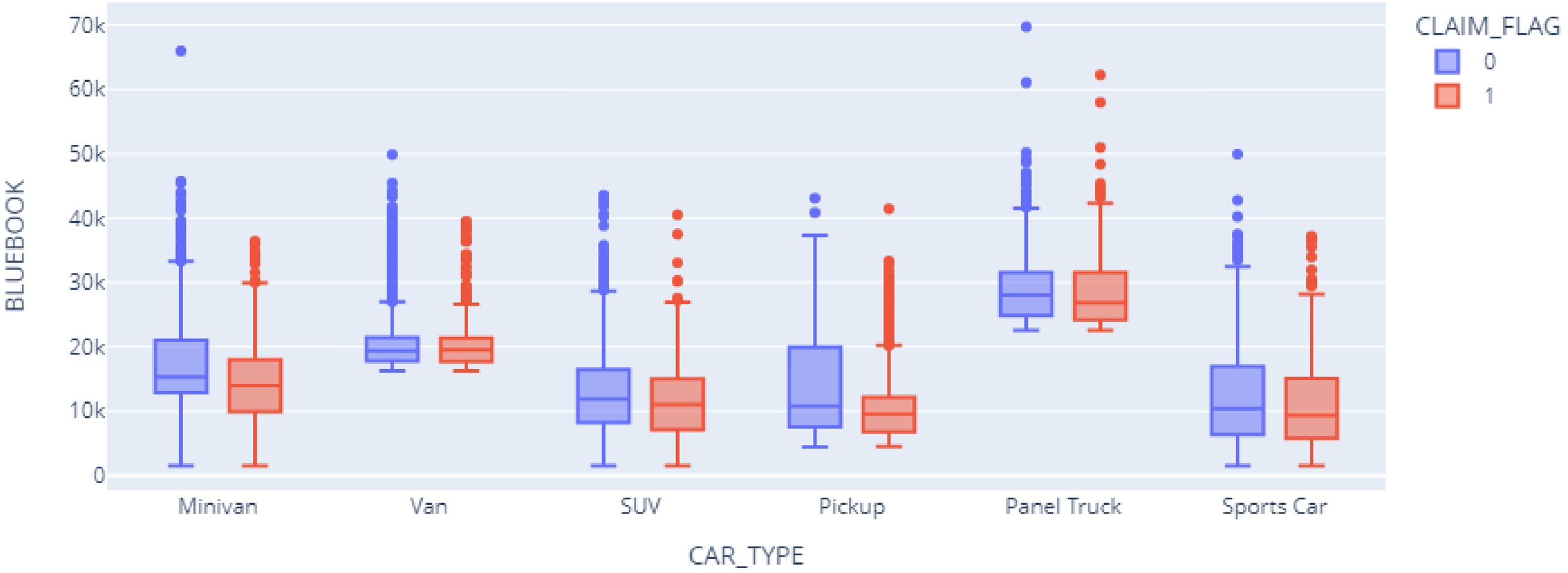
VALUE OF CAR

Box Plot about the value of car in each occupation



VALUE OF CAR

Box Plot about the value of car in each occupation



MACHINE LEARNING

CONTENT

HANDLE IMBALANCE DATA

FEATURE ENGINEERING

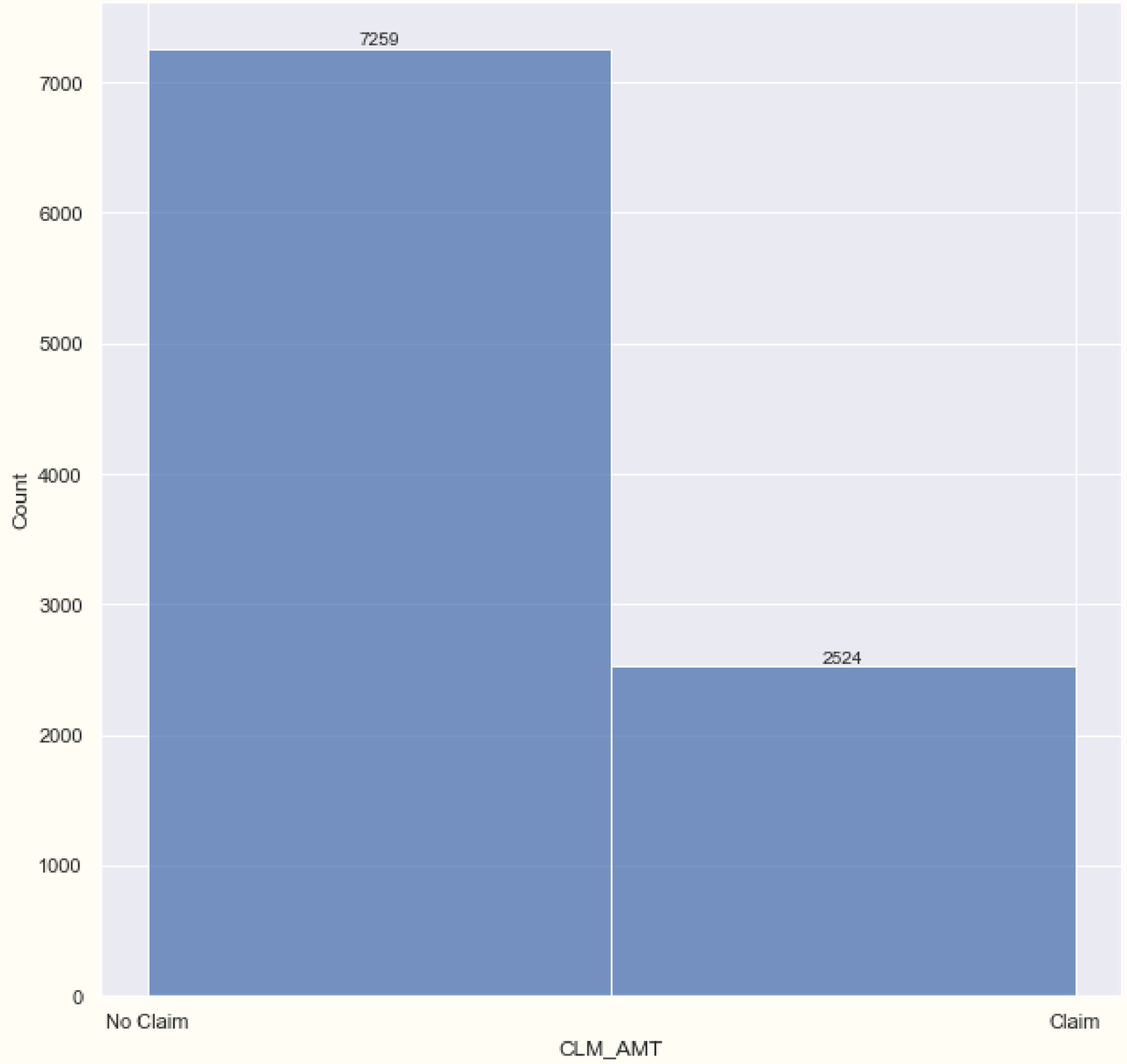
MODEL TRAINING

FEATURE ENGINEERING

2-VALUE CATEGORY

3 OR MORE VALUES CATEGORY

HANDLE IMBALANCE DATA

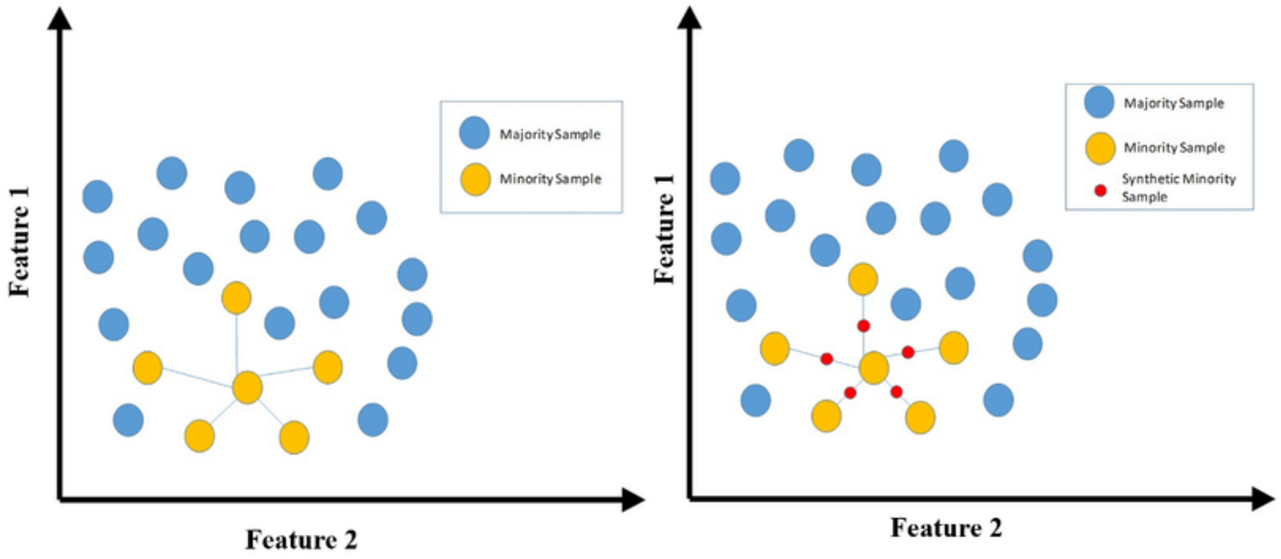


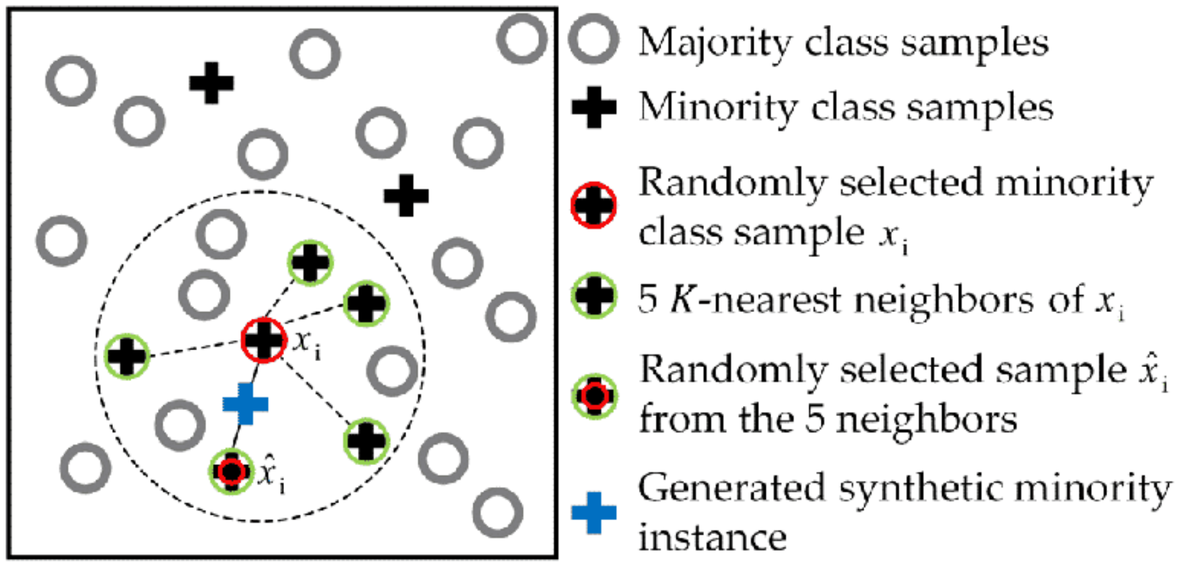
74.2%

THE RATE OF CUSTOMERS WHO WILL NOT USE CAR INSURANCE IN THE FUTURE

HANDLE IMBALANCE DATA

SMOTE





ALGORITHM

XGBOOST

LOGISTIC REGRESSION

RANDOM FOREST

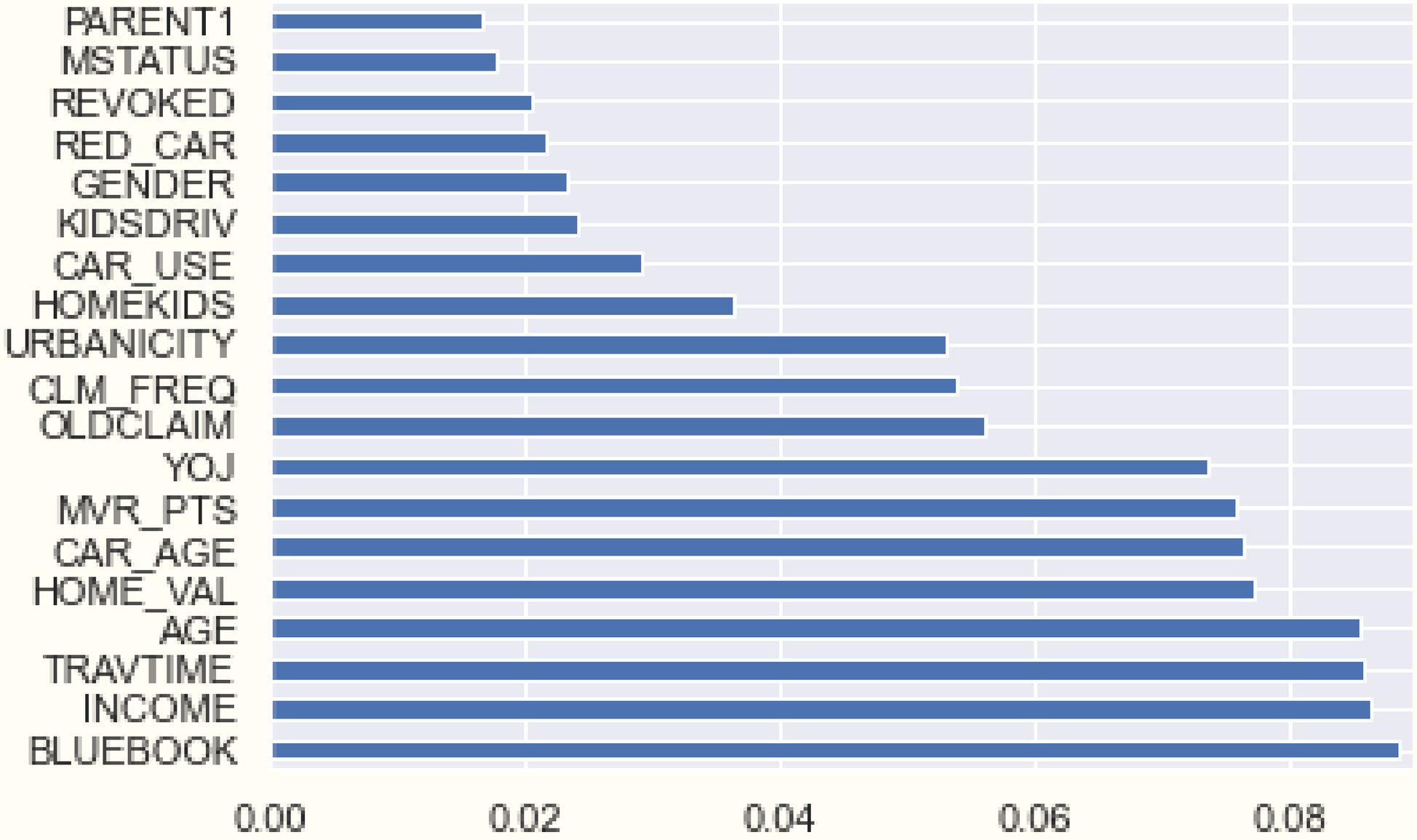
ALGORITHM

ADA BOOST

DECISION TREE

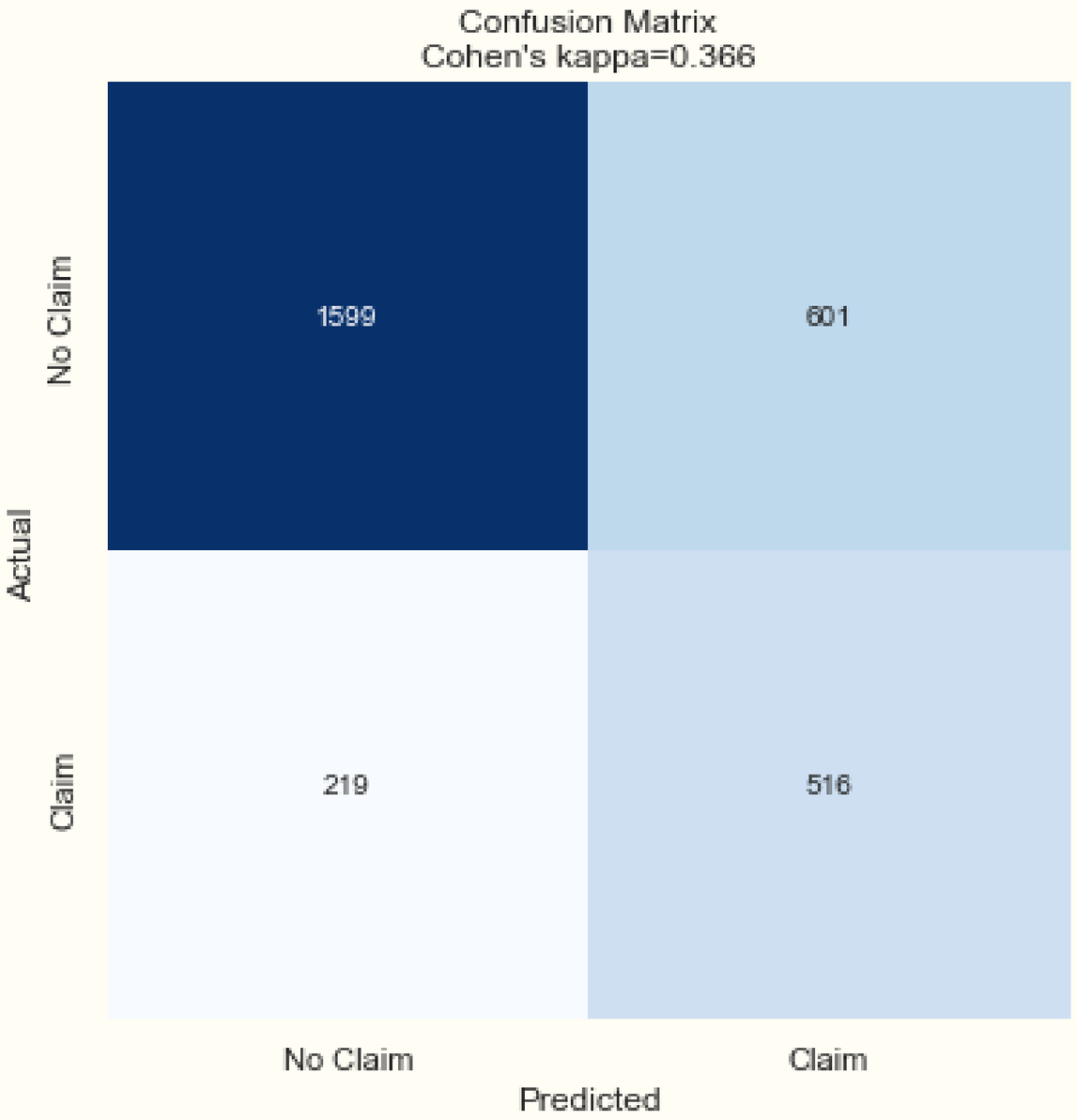
GRADIENT BOOSTING

FEATURE SELECTION



BLUEBOOK
INCOME
TRAVTIME
AGE
HOME_VAL

LOGISTIC REGRESSION



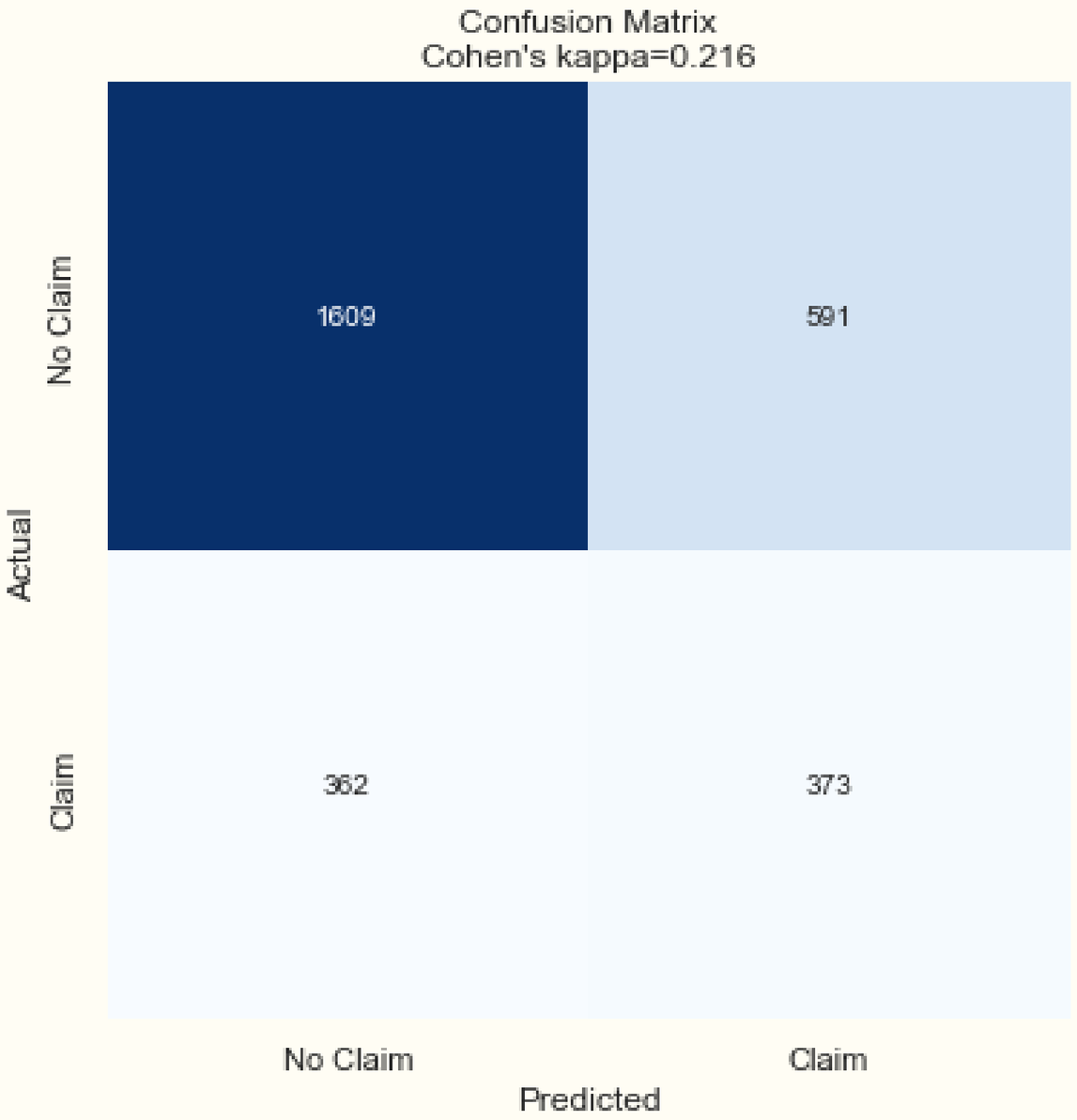
ACCURACY: 0.721

PRECISION: 0.462

RECALL: 0.702

F1 SCORE: 0.557

DECISION TREE



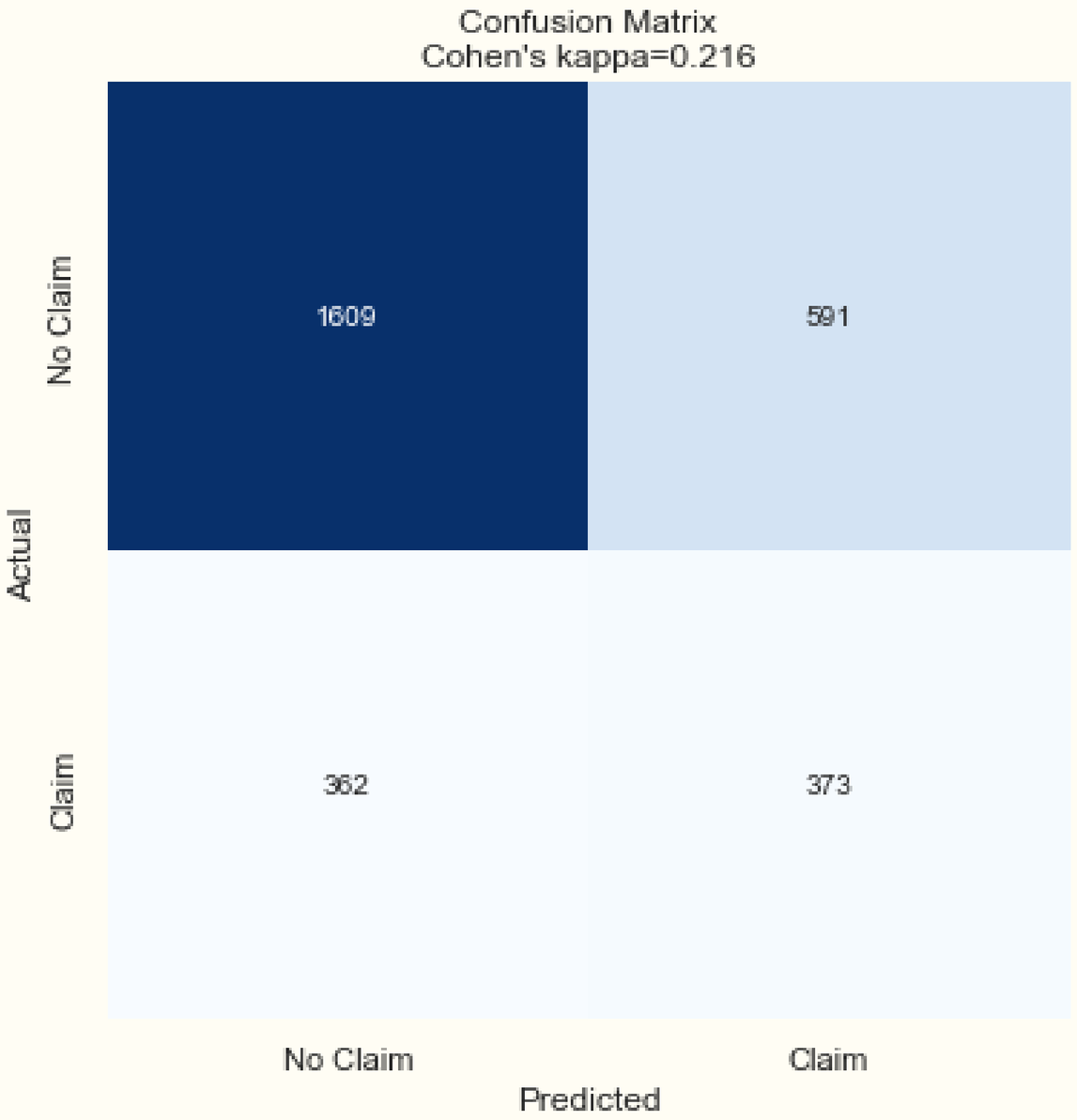
ACCURACY: 0.675

PRECISION: 0.387

RECALL: 0.507

F1 SCORE: 0.439

XGBOOST



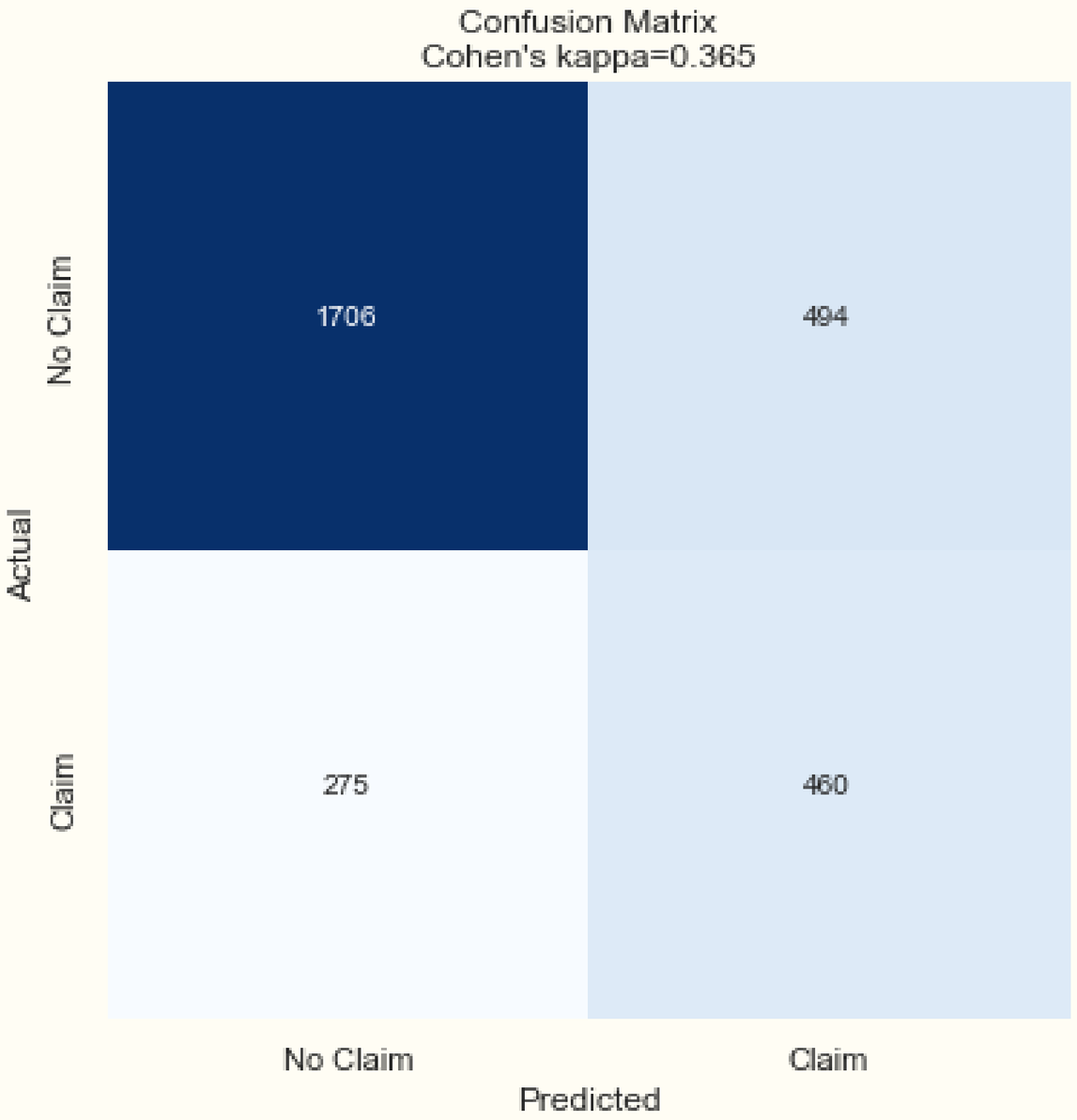
ACCURACY: 0.734

PRECISION: 0.478

RECALL: 0.653

F1 SCORE: 0.552

RANDOM FOREST



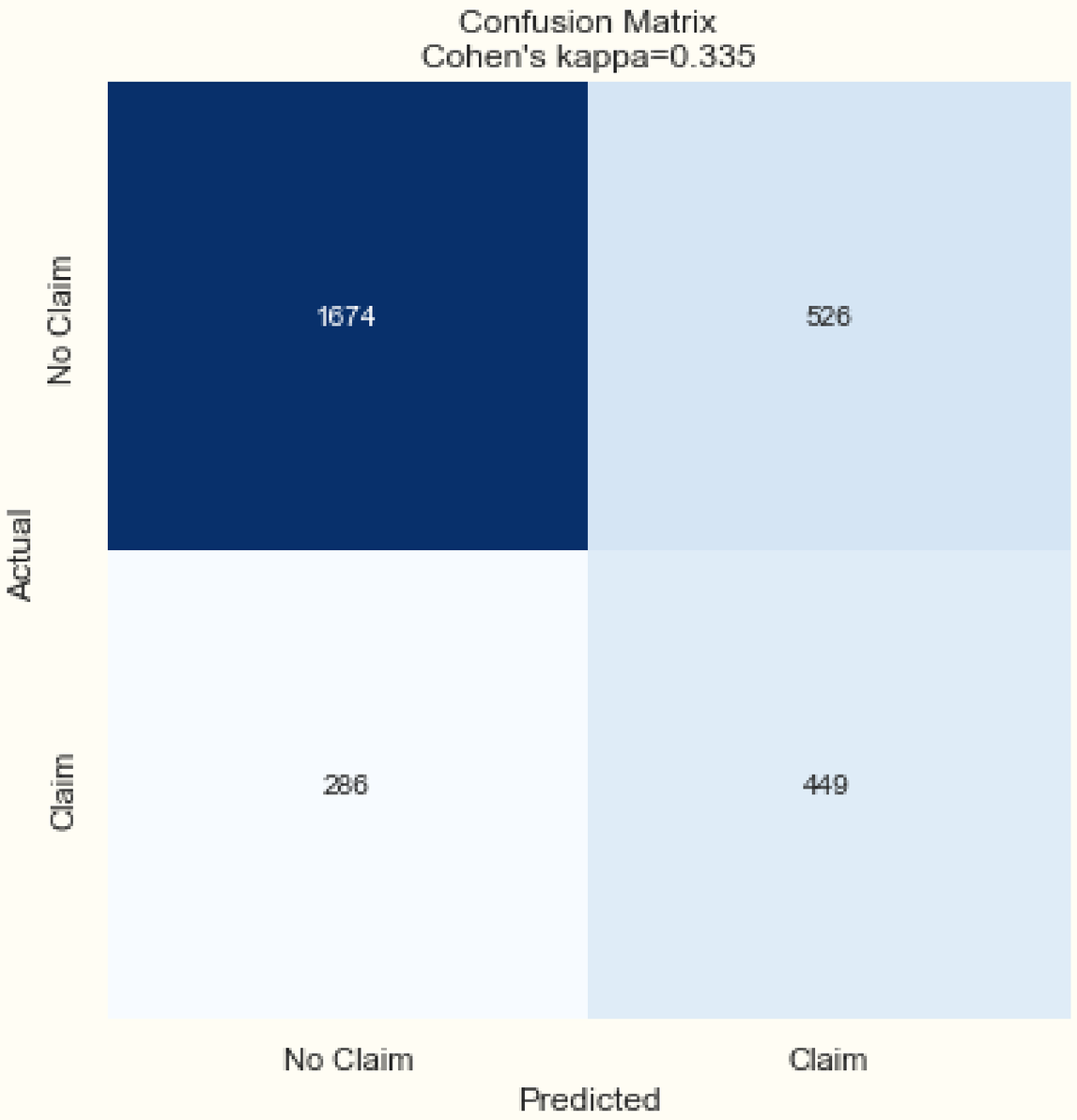
ACCURACY: 0.738

PRECISION: 0.482

RECALL: 0.626

F1 SCORE: 0.545

ADA BOOST



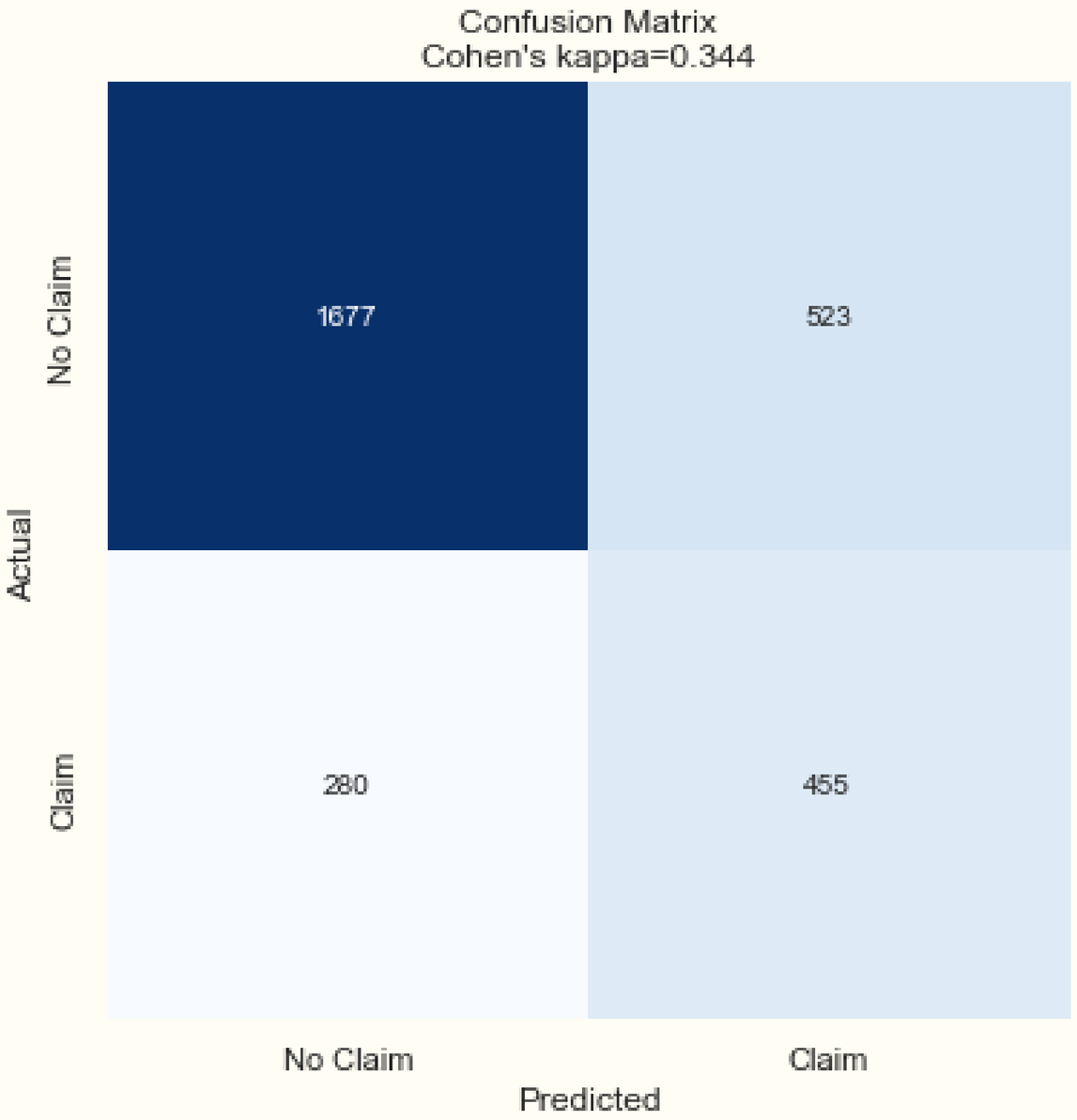
ACCURACY: 0.723

PRECISION: 0.461

RECALL: 0.611

F1 SCORE: 0.525

GRADIENT BOOSTING



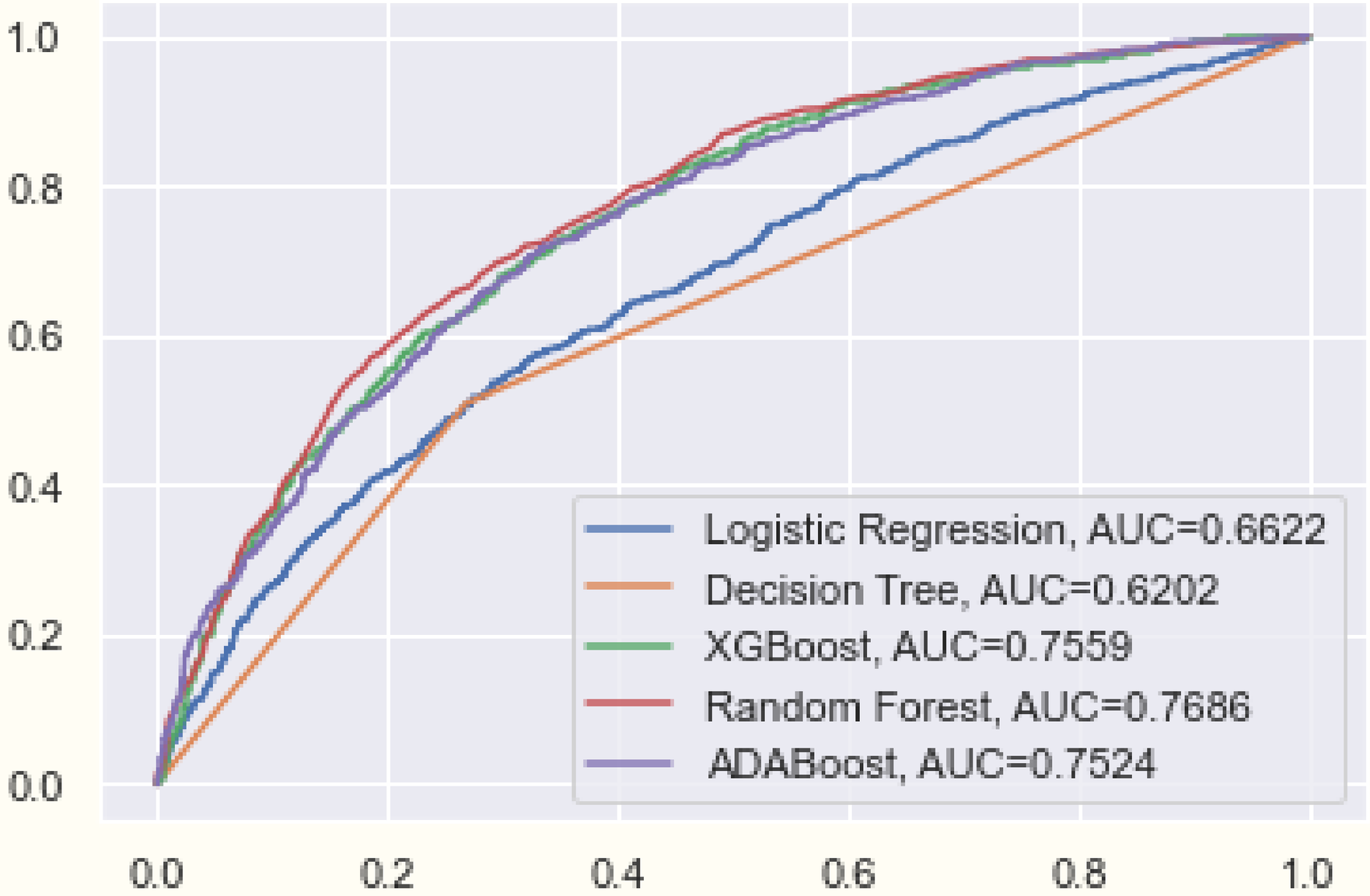
ACCURACY: 0.726

PRECISION: 0.465

RECALL: 0.619

F1 SCORE: 0.531

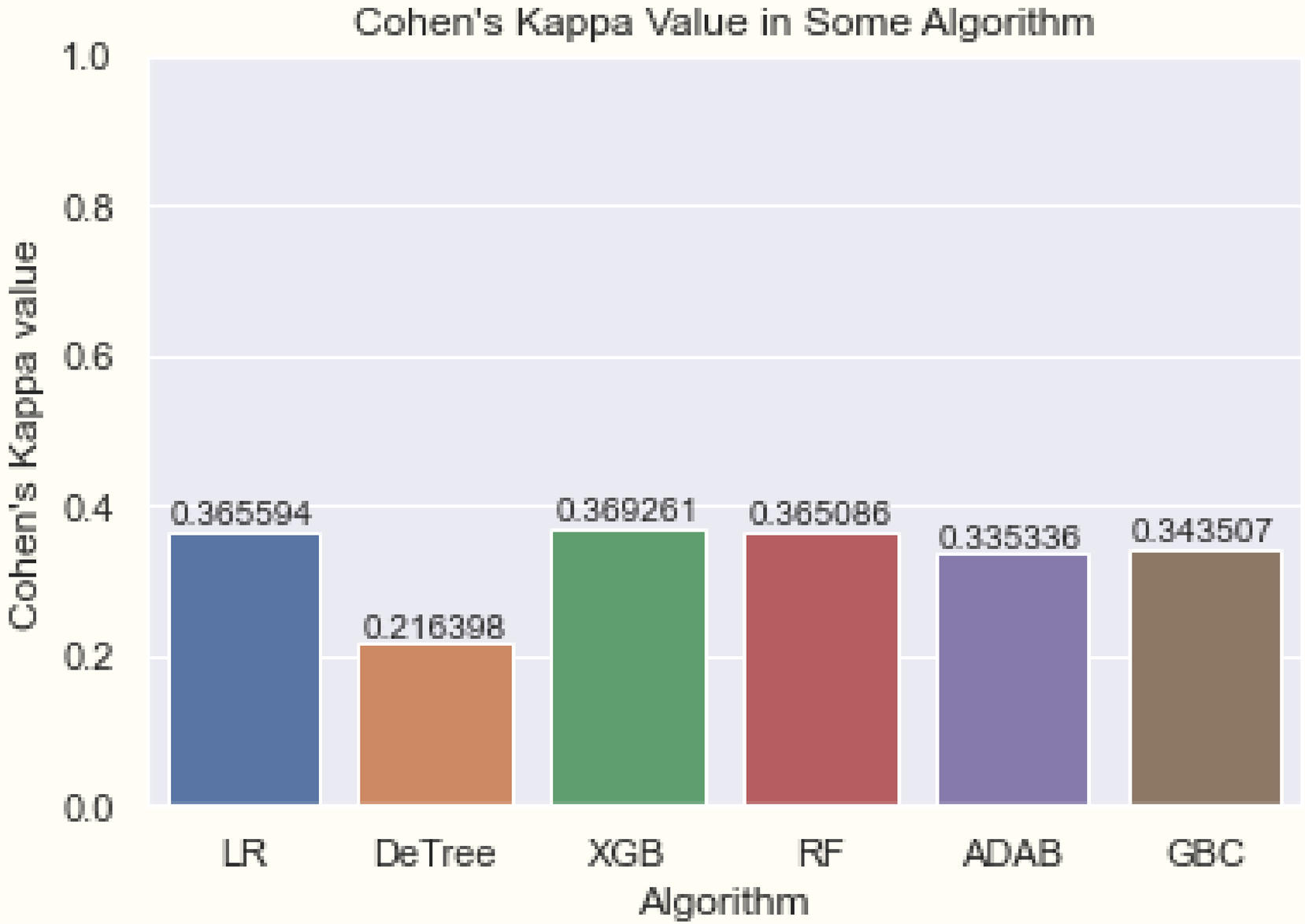
AUC AND ROC



0.7686

THE AUC SCORE OF RANDOM FOREST ALGORITHM

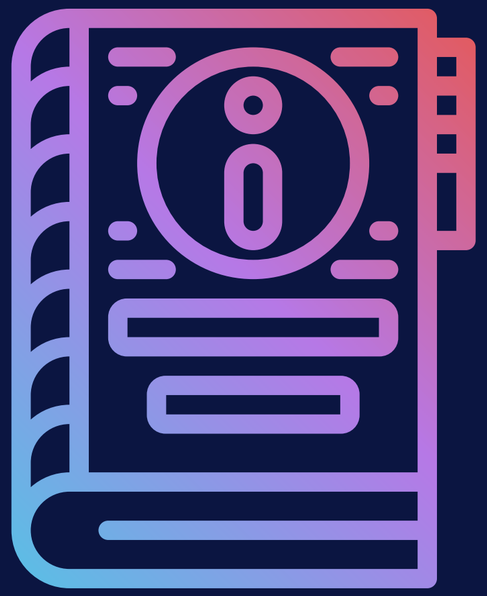
COHEN'S KAPPA



0.369

COHEN'S KAPPA SCORE OF XGBOOST ALGORITHM

REFERENCES



REFERENCE

- Abdelhadi, S., Elbahnasy, K., & Abdelsalam, M. (2020). A proposed model to predict auto insurance claims using machine learning techniques. Journal of Theoretical and Applied Information Technology, 98(22).
- Batista, G. E., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence, 17(5–6), 519–533.
- Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. Pattern Recognition, 74, 406–421.
- Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. Journal of Big Data, 8(1), 1–37.

REFERENCE

- Hyperparameter optimization. Adaboost. (2019). Kaggle. https://www.kaggle.com/code/juanmah/tactic-03-hyperparameter-optimization-adaboost
- Jain, A. (2016). Mastering XGBoost Parameter Tuning: A Complete Guide with Python Codes. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/
- Khusna, W., & Murfi, H. (2020). An analysis of the proportion of feature subsampling on XGBoost-A case study of claim prediction in car insurance. AIP Conference Proceedings, 2296(1), 20058.
- Li, X. (2023). Identifying the Optimal Machine Learning Model for Predicting Car Insurance Claims: A Comparative Study Utilising Advanced Techniques. Academic Journal of Business & Management, 5(3), 112–120.

REFERENCE

- Logistic Regression - hyperparameter tuning. (2021). Kaggle. https://www.kaggle.com/code/funxexcel/p2-logistic-regression-hyperparameter-tuning
- Neumann, Ł., Nowak, R. M., Okuniewski, R., & Wawrzyński, P. (2019). Machine Learning-Based Predictions of Customers' Decisions in Car Insurance. Applied Artificial Intelligence, 33(9), 817–828.
- Qiao, F. (2019). Logistic Regression Model Tuning with scikit-learn — Part 1. Toward Data Science. https://towardsdatascience.com/logistic-regression-model-tuning-with-scikit-learn-part-1-425142e01af5
- Satpathy, S. (2020). Overcoming Class Imbalance using SMOTE Techniques. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

THE END