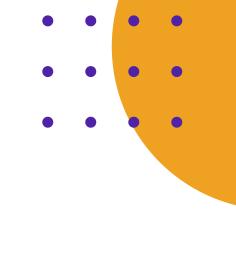


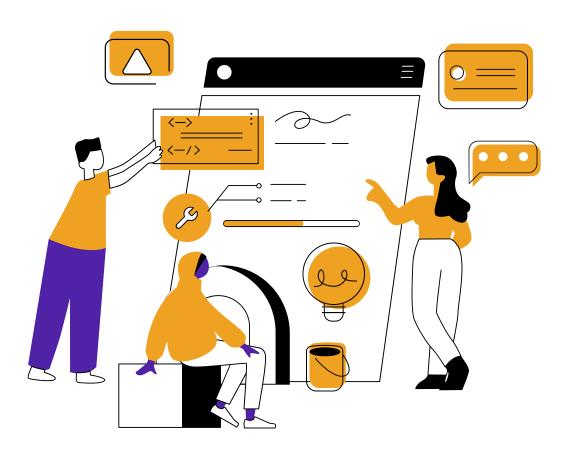
Health Insurance

Cross Sale Prediction

Student: Nguyen Anh Tai - FX13245

Mentor: Nguyen Huy Thanh







The presentation would go through the following main points



I. PROBLEM STATEMENT

II. OBJECTIVES

III. METRICS

IV. DATA UNDERSTANDING

V. EXPLORATORY DATA ANALYSIS

VI. DATA PREPROCESSING

VII. FEATURE SELECTION

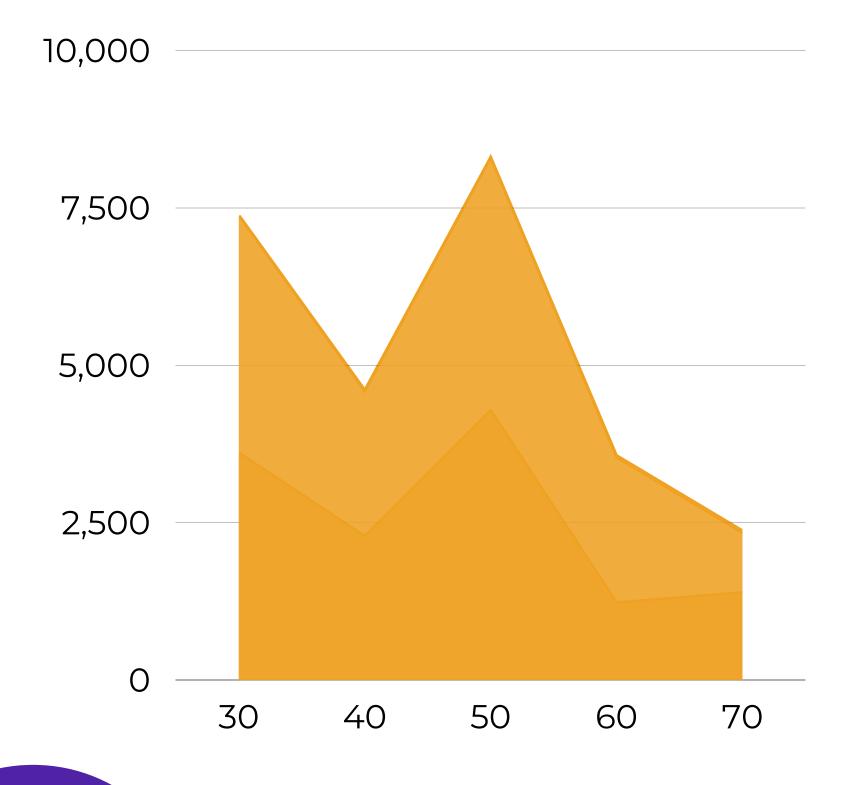
VIII. BUILD MODEL

IX. IMPROVEMENT



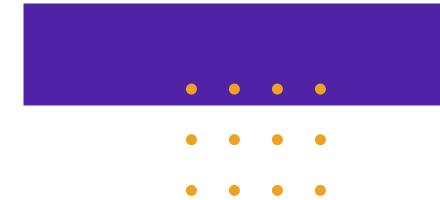
Problem Statement

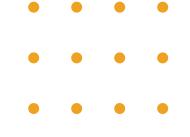
The client in the context is an insurance company which has provided health insurance to the customers. Now, they need a model which helps predict the likelihood that the customers from the previous year will choose the vehicle insurance offered by the company.

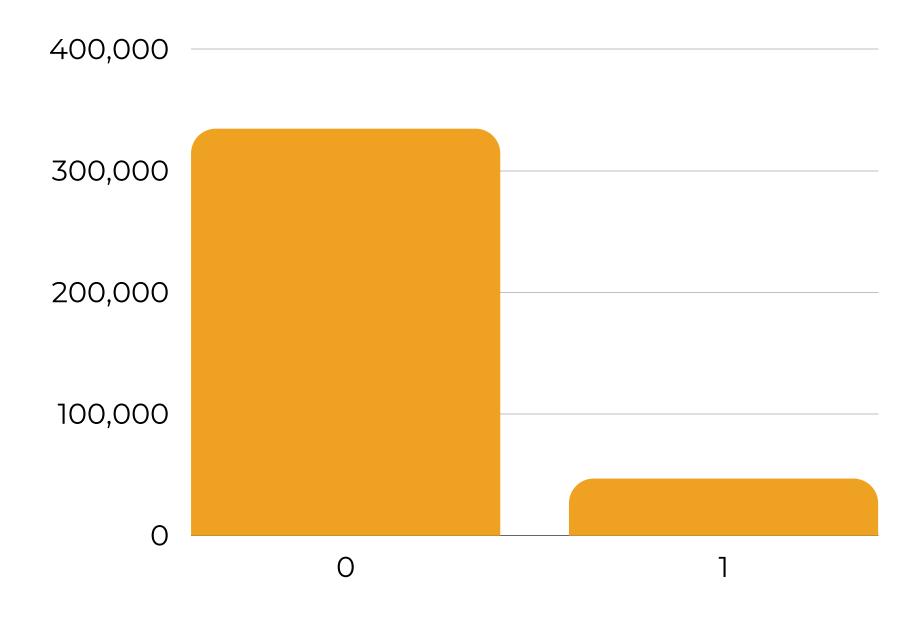


Objectives

- Identify the driving forces behind the conversion to vehicle insurance among the customers
- Build a model which predicts the potential customers who would be interested in vehicle insurance

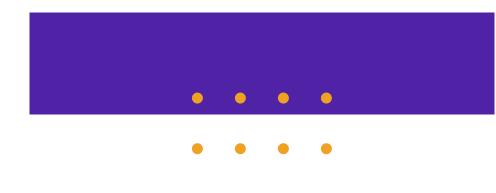






Metrics

- A good model should obtain high f1 score (at least 0.4)
- For the purpose of profit generation, high recall rate (at least 0.8) is preferable



- ID: Unique ID for the customers
- Gender: Gender of the customers
- Age: Age of the customers
- Driving license: 0 no DL, 1 Already have
- Region Code: Unique code for the customers' region
- Previously Insured: 0 no vehicle insurance, 1 -Already have
- Vehicle Age: Age of the vehicle
- Vehicle Damage: 0 not get vehicle damaged in the past, 1 - get vehicle damaged in the past
- Annual Premium: The amount of premium paid in a year
- Policy Sales Channel: Channel of outreaching to the customers
- Vintage: Number of days associated with the company
- Response: 0 customer not interested, 1 customer is interested

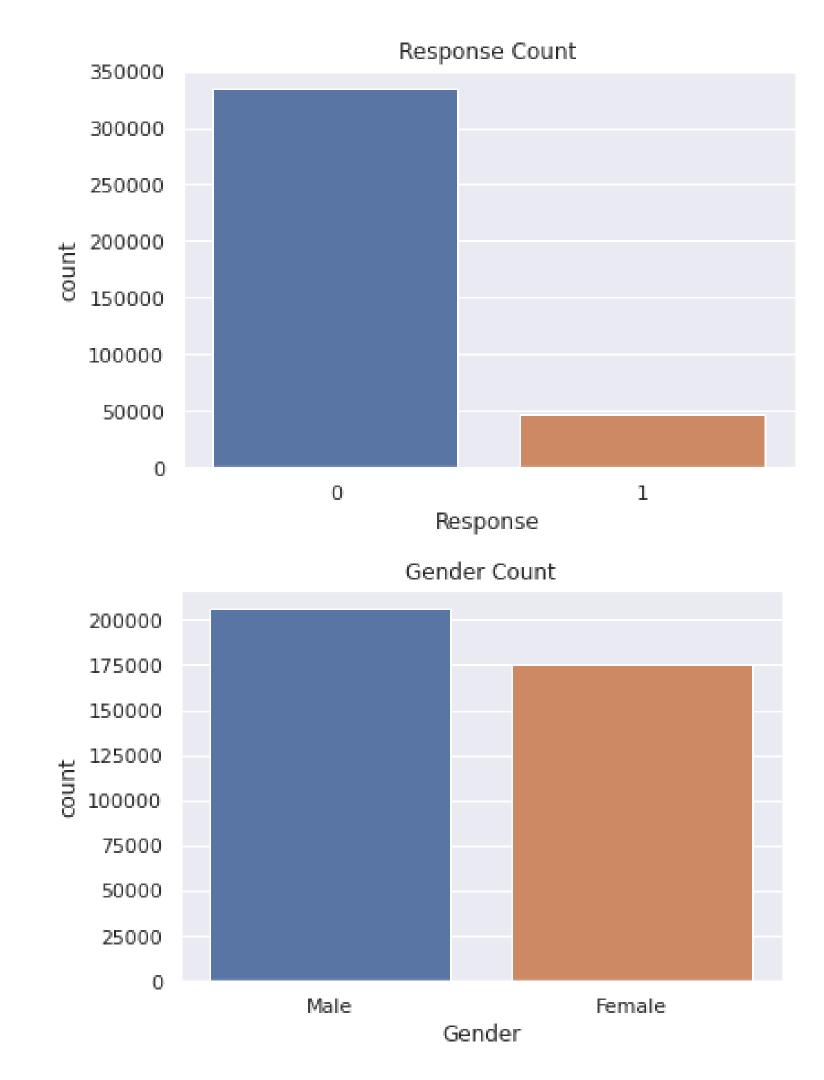
Data Understanding





Dataset is **imbalanced** towards the **negative** target (**uninterested** responses)

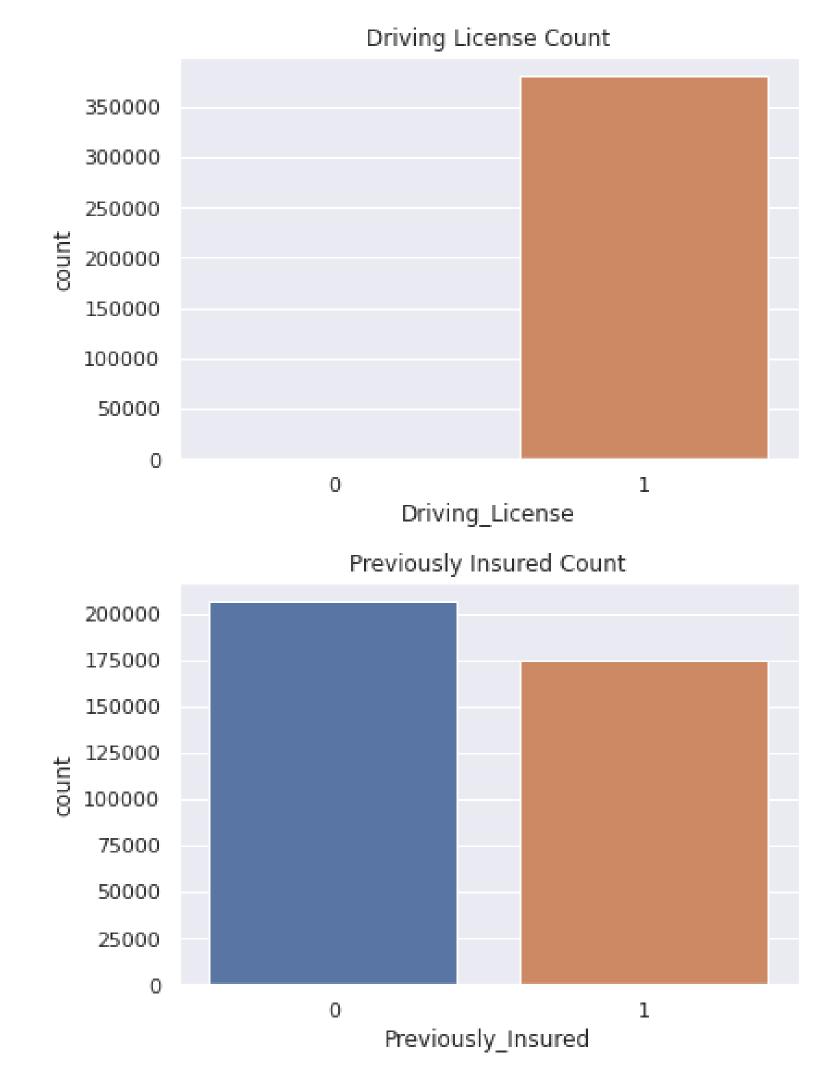
Dataset has a **balanced gender** distribution





A great majority of customers have **driving license**

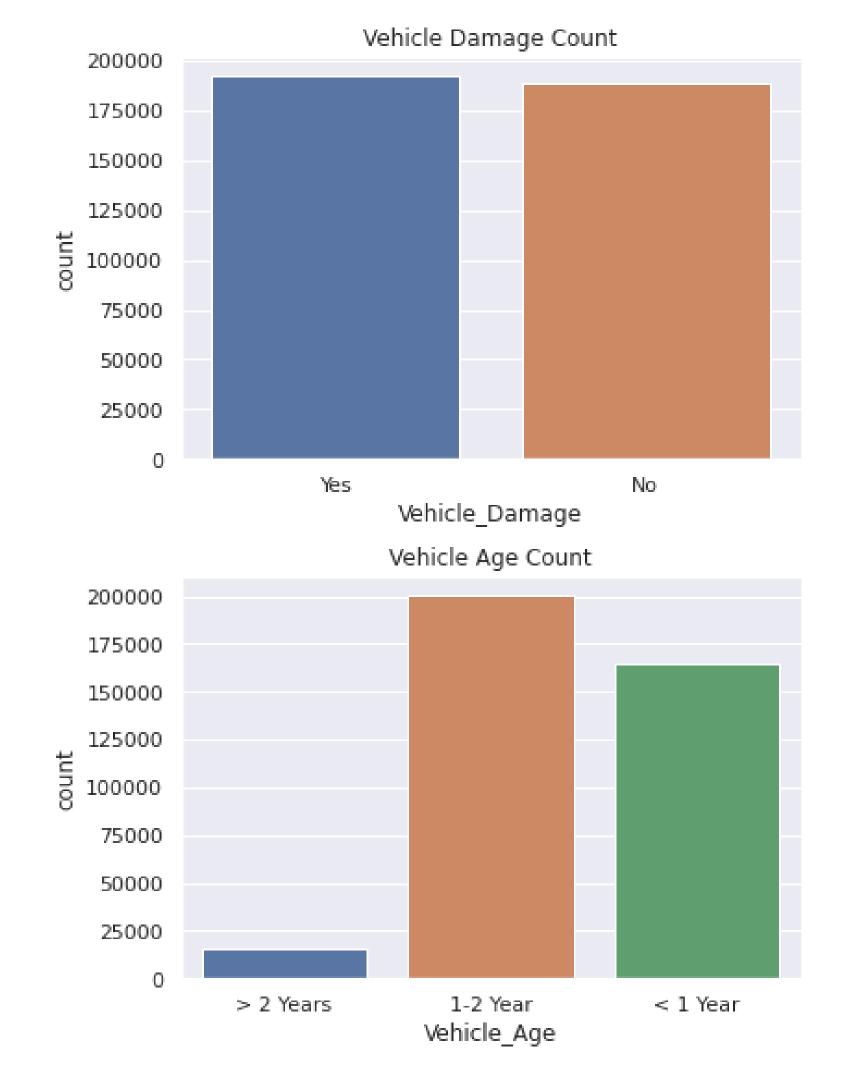
A great number of customers are **not** previously insured --> potential customers





The dataset is **balanced** regarding the number of customers previously **having vehicle damage**

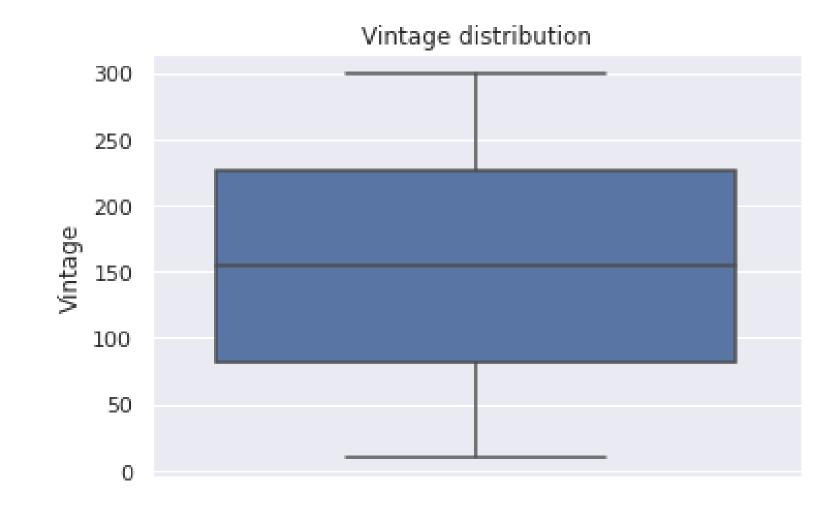
The dataset is **biased** towards the customers owning a new car (**fewer than 2 years**) --> not generalize for customers owing a car **more than 2 years**

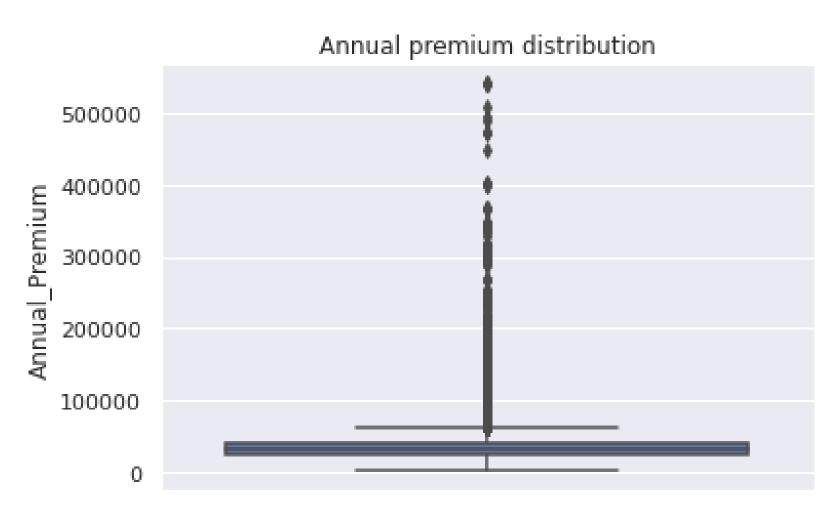




The feature vintage appears to have no outliers with a wide range of associated days

The annual premium has outliers and most of the customers pay less than 100000 ruppee

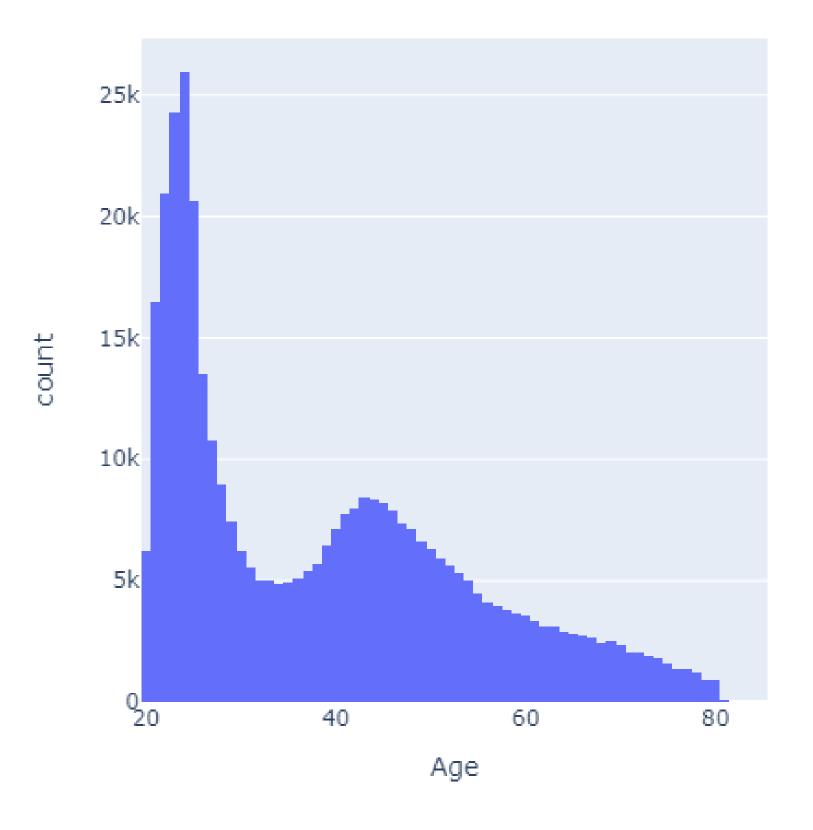






The dataset contains mostly the customers at a young age (younger than 35 years old)

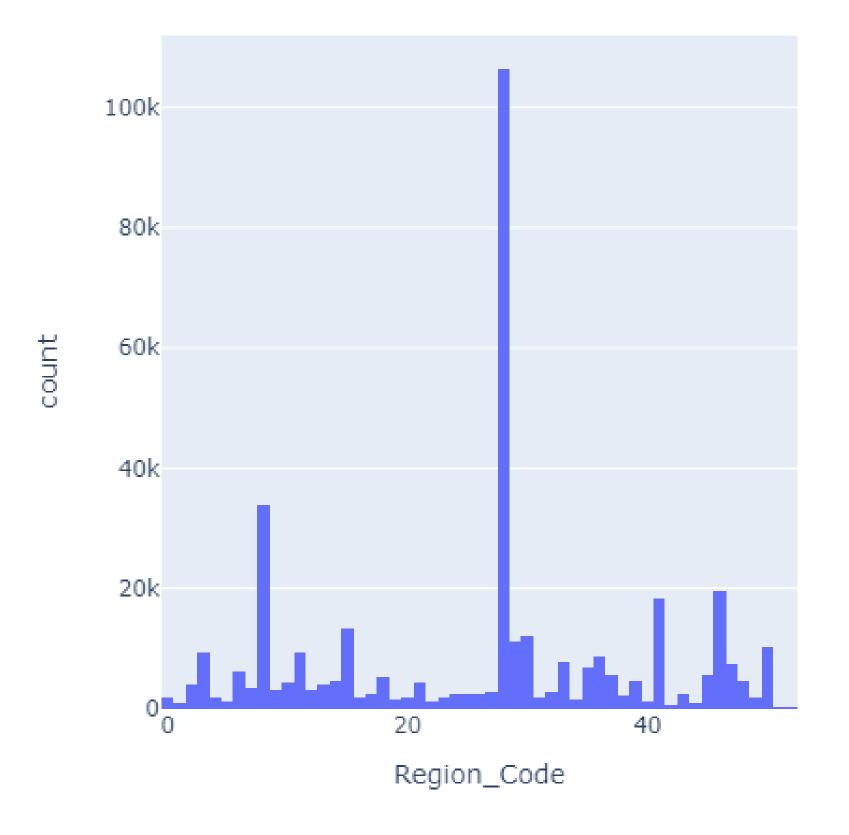
Age distribution





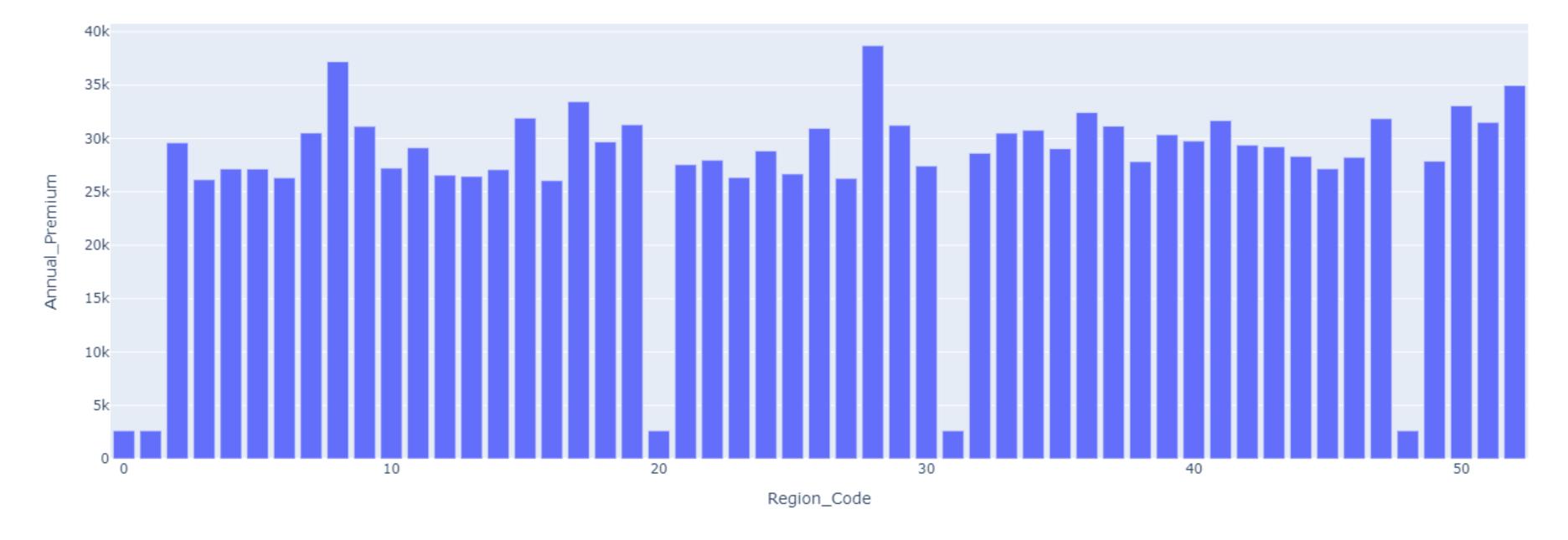
The majority of the customers come from region code 8 and 28

Region_Code distribution





Region code **8 and 28** also have the *highest median/average annual premium*

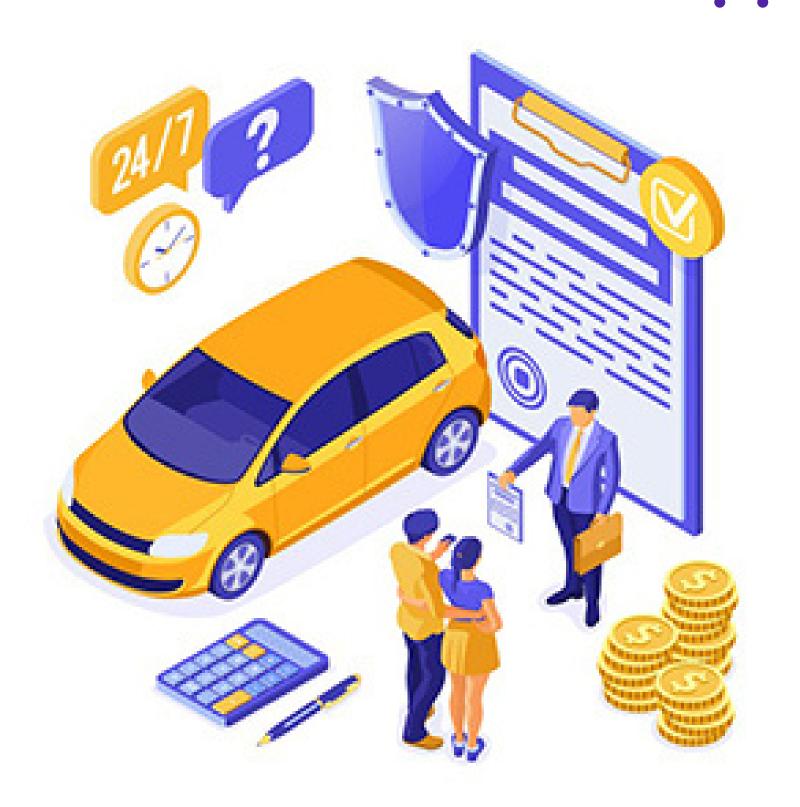


Key Findings:

- The dataset is highly imbalanced, biased towards the uninterested responses
- Nearly all the customers in the dataset have a driving license
- Majority of the customers were not previously insured for vehicle damage
- All the customers seem to be the new vehicle owner with less than 2 years of ownership
- The customers seem to pay very low for annual premium
- The customers come primarily from the young
 age group and the region code 8, 28

Exploratory Data Analysis

How to increase conversion rate for vehicle insurance?



Targeted points

Identify the channels which mostly attract the potential customers

Identify the regions that should be targeted at

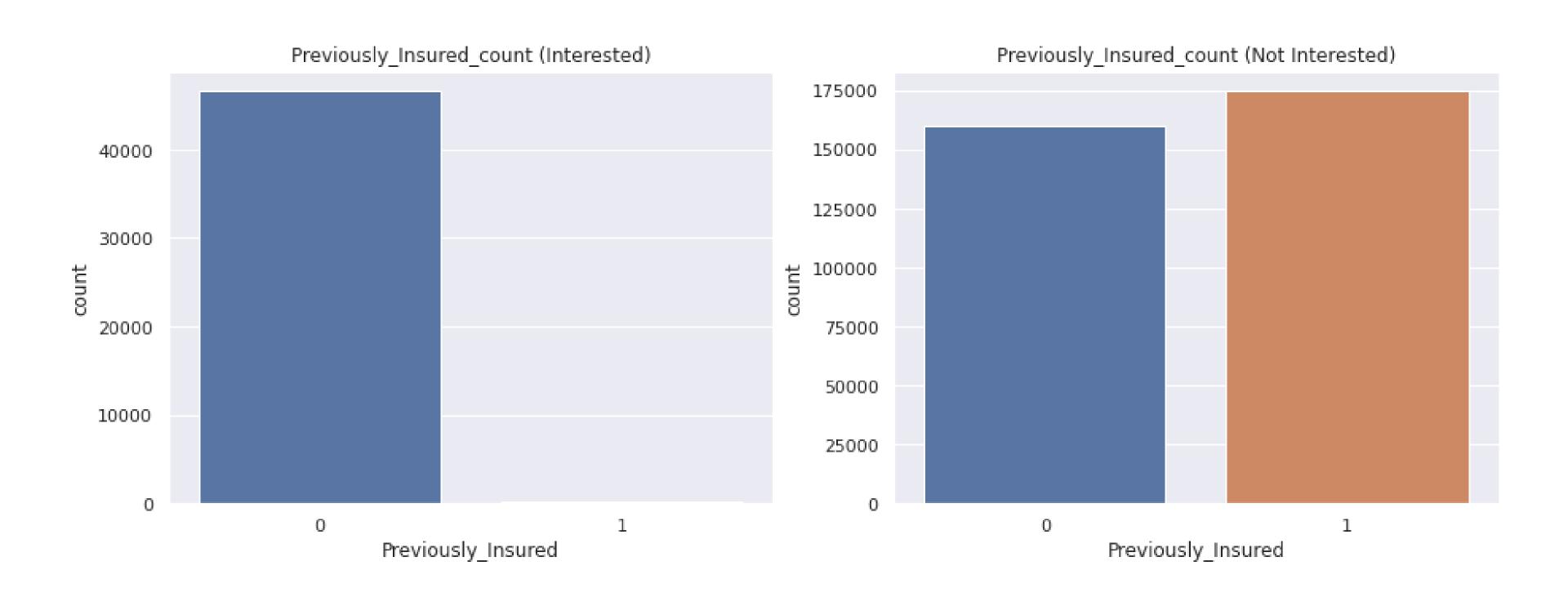
ı

How to increase the sales of vehicle insurance?

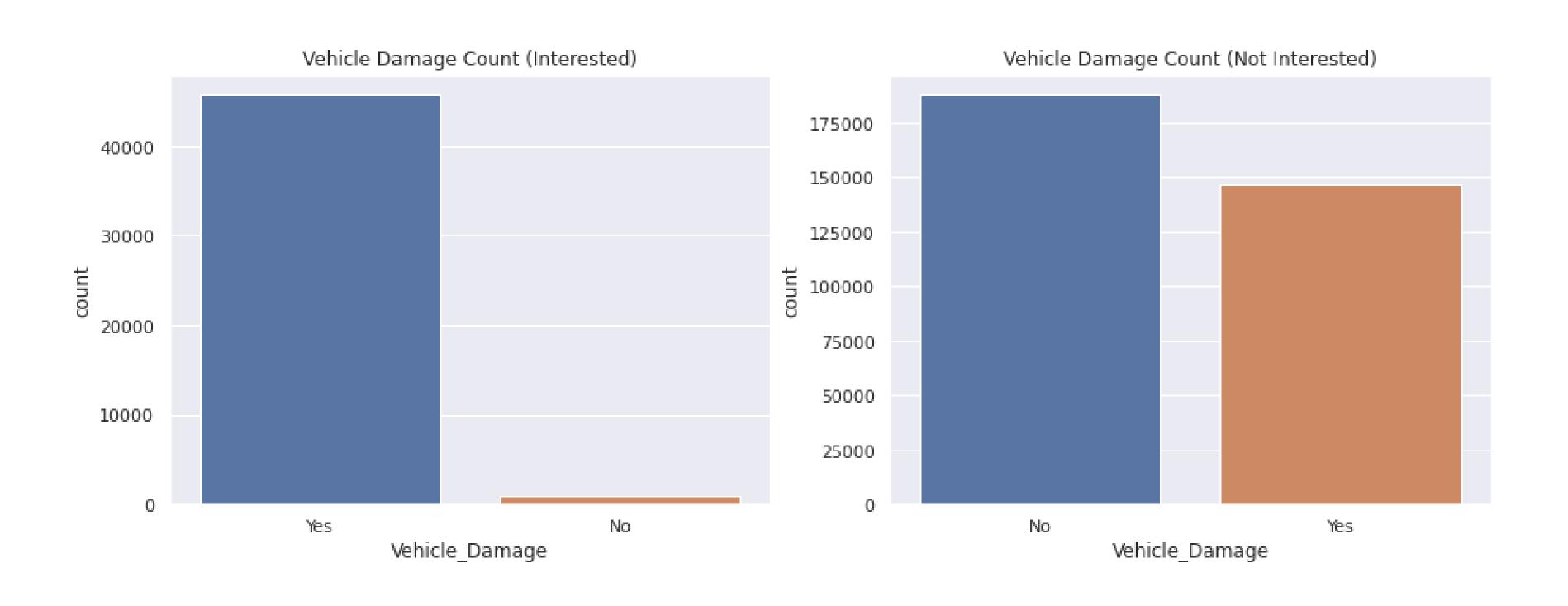
Identify the similar traits among the potential customers

Previous insurance Age group Vehicle damage Gender Vehicle age Age group Gender of customers Influence of age groups Vintage

Attention to the previously insured customers?

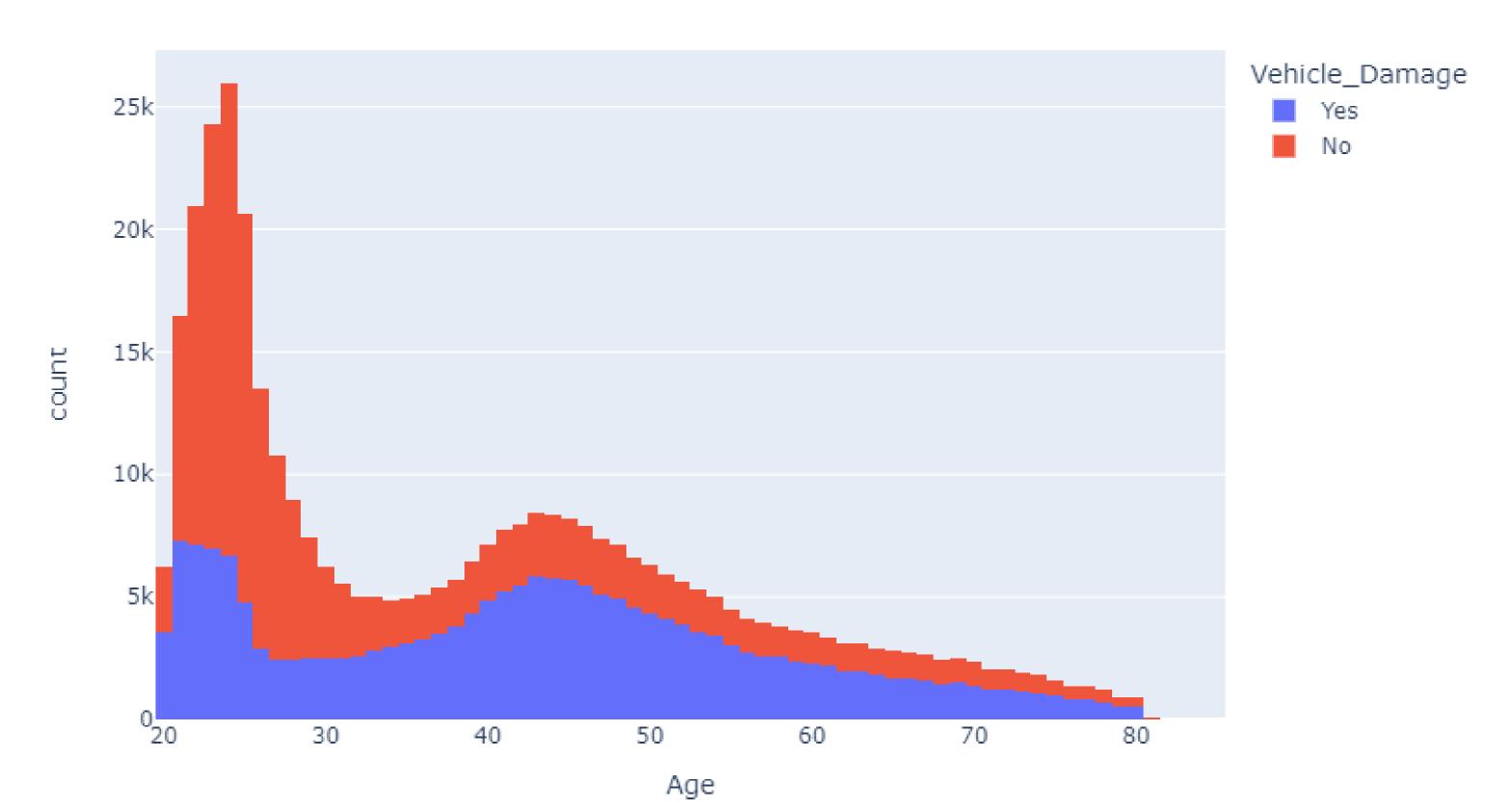


Attention to the customers previously having vehicle damage?

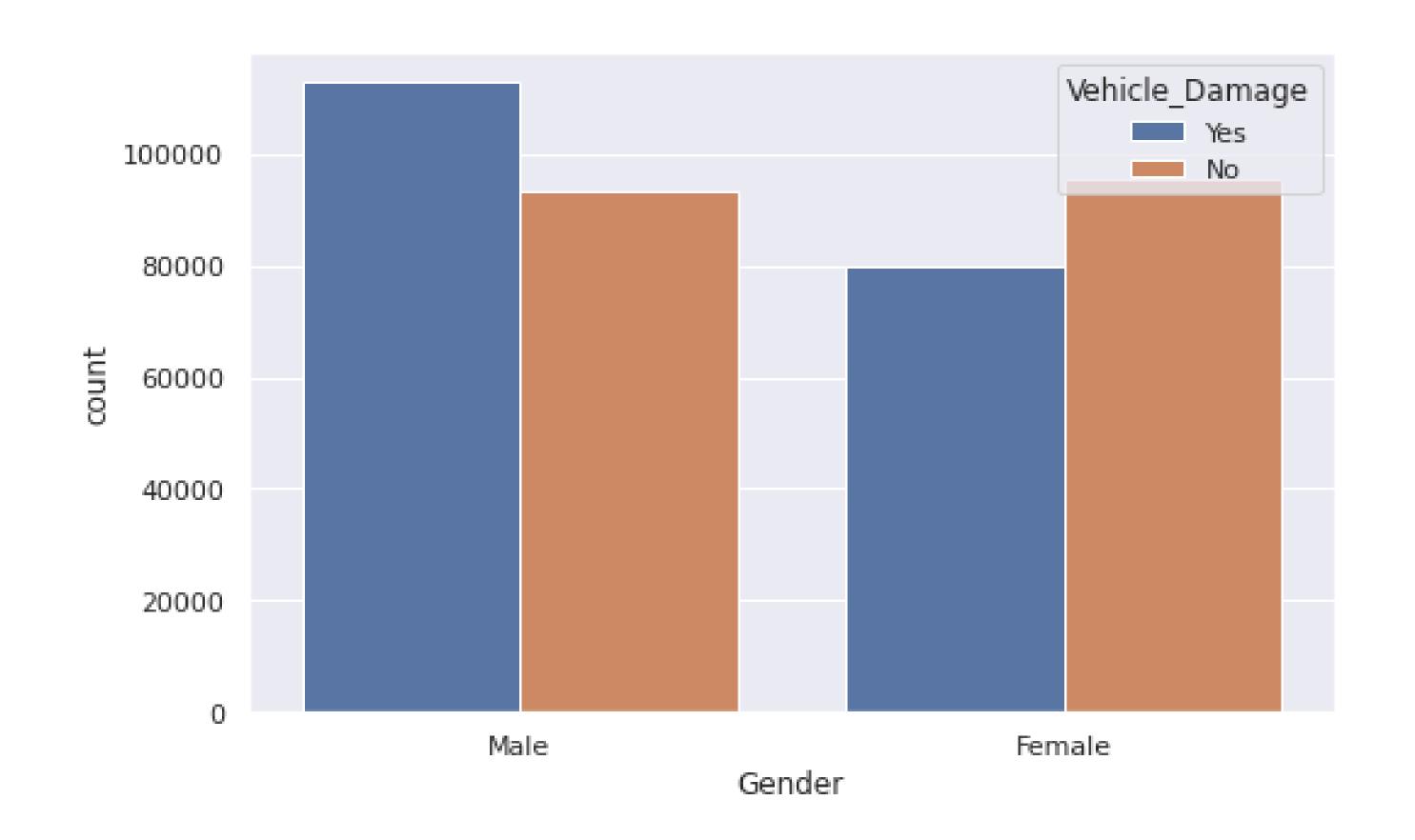


Who tends to be involved in vehicle damage and at which age?

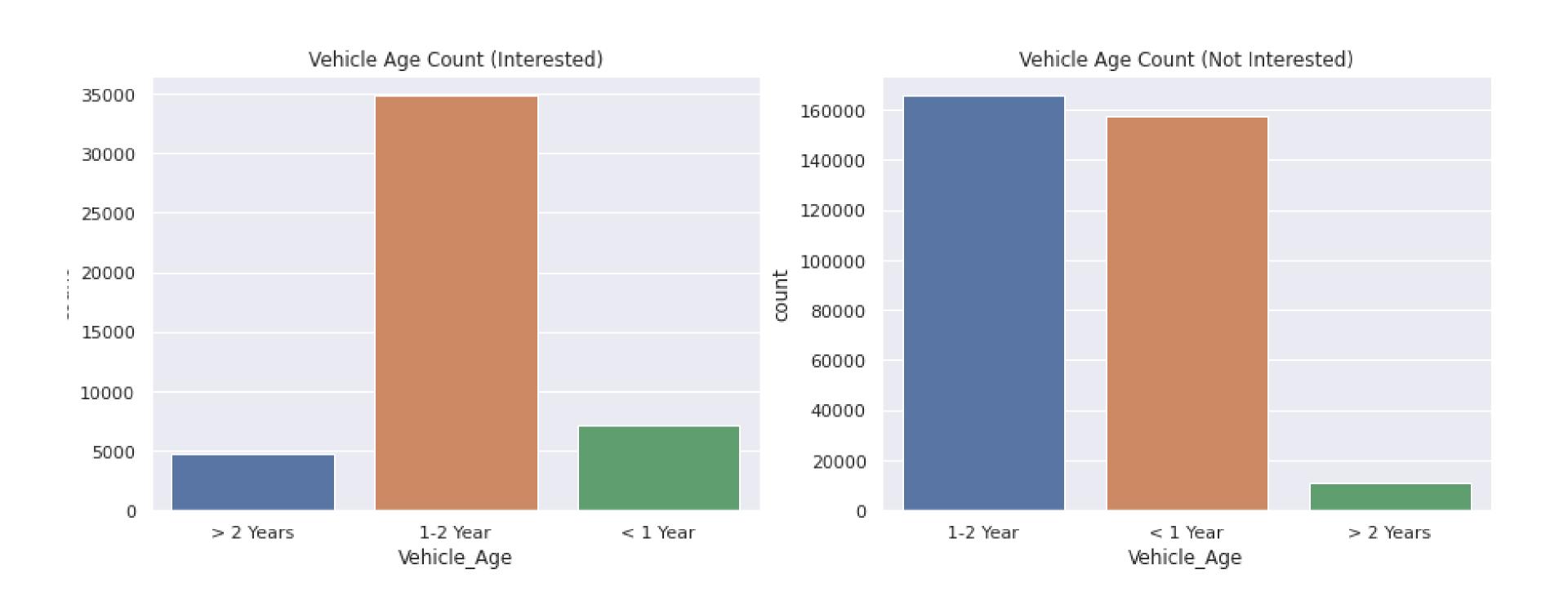
Vehicle damage for different ages



Does gender influence the likelihood of vehicle damafe?

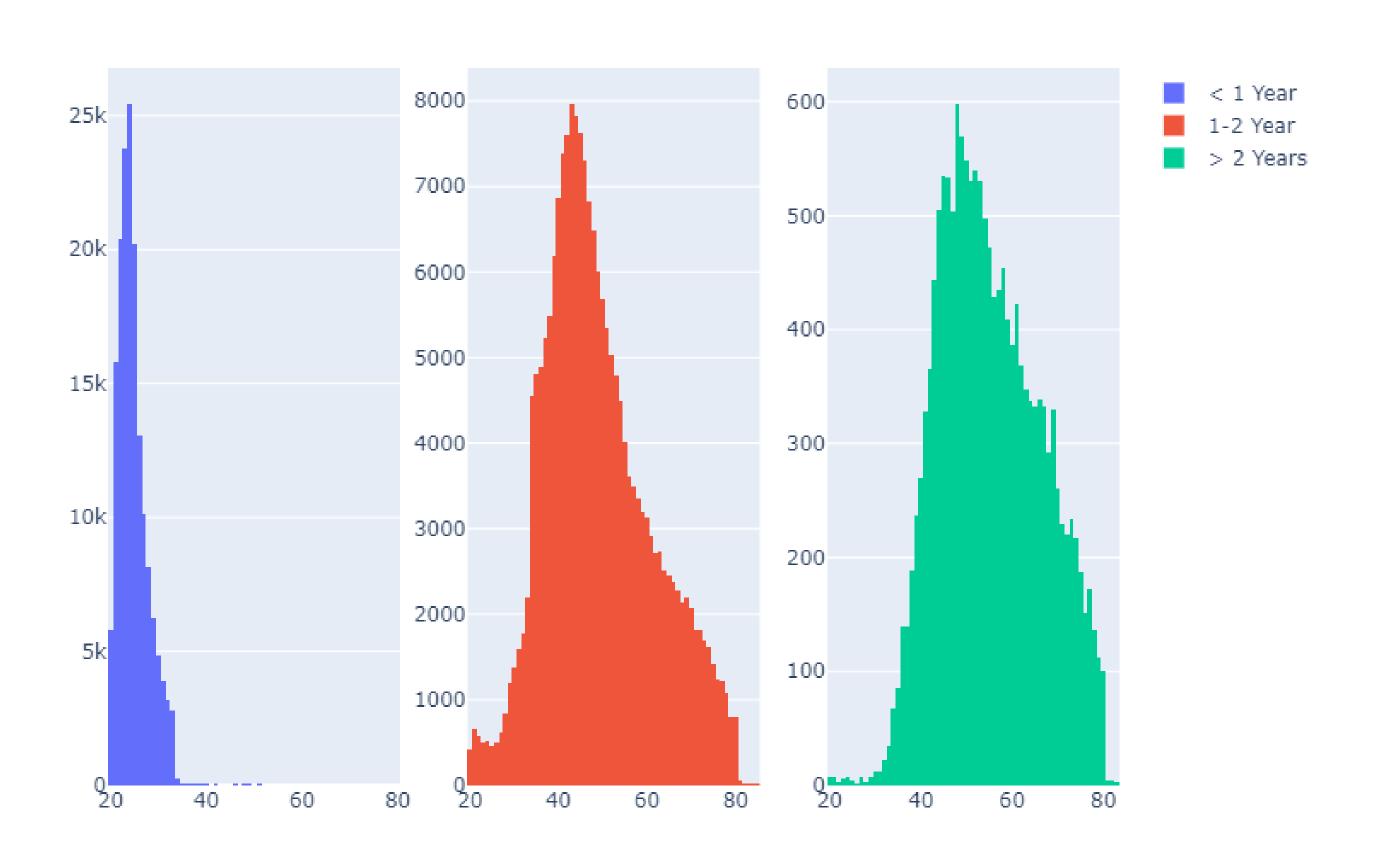


Does vehicle age influence the decision of the customers to buy vehicle insurance?

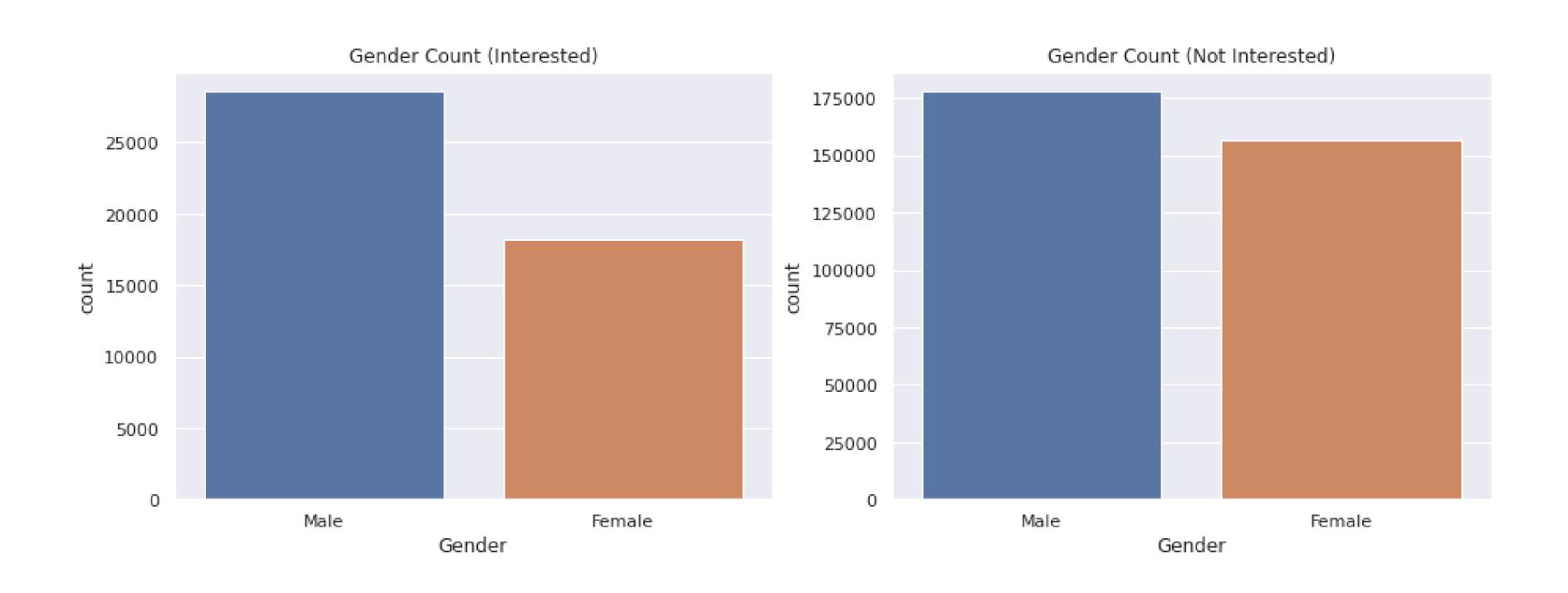


Does age correlate with the vehicle age among the customers?

Vehicle Age and Customer's age

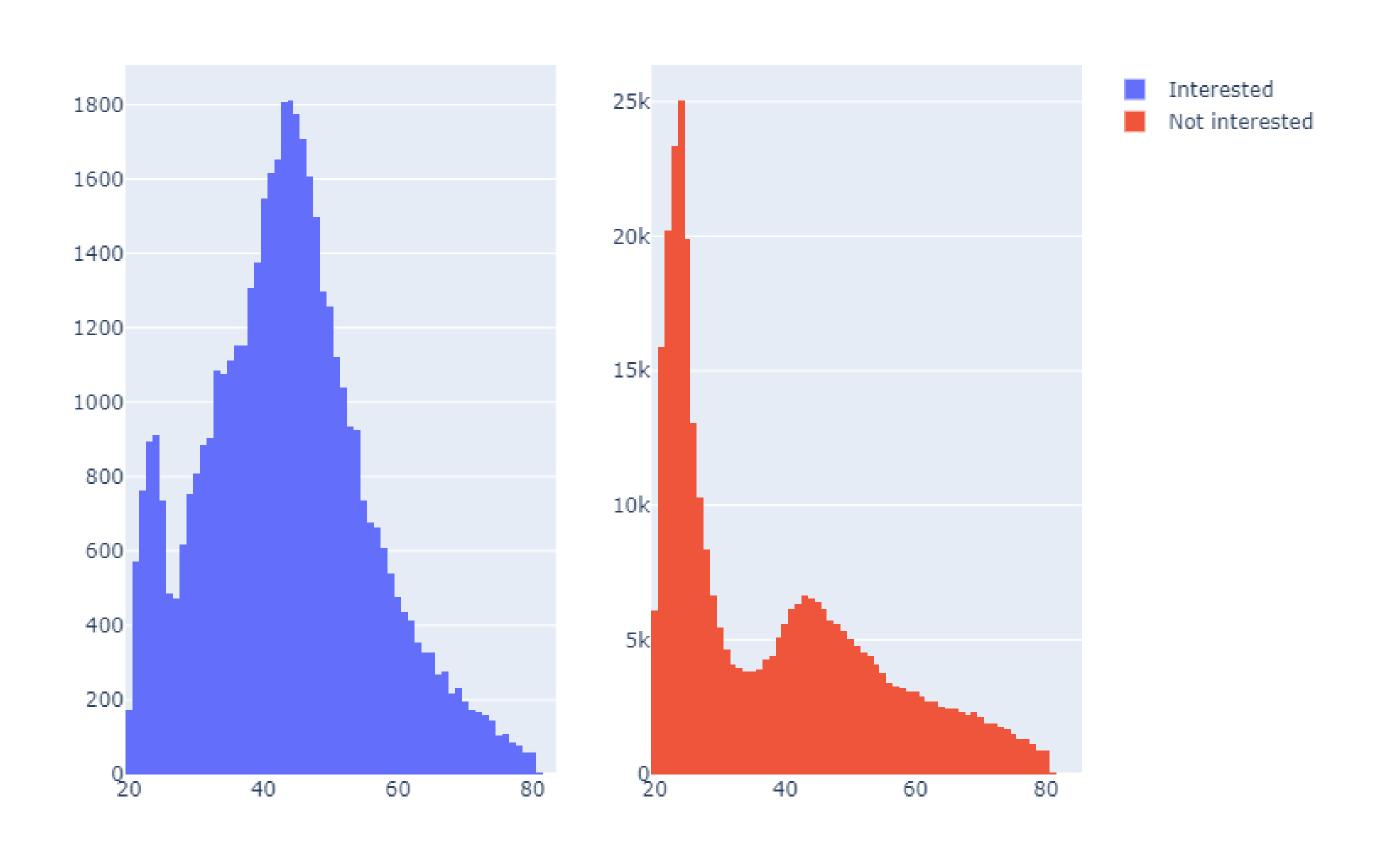


Does gender influence the decision to buy vehicle insurance?



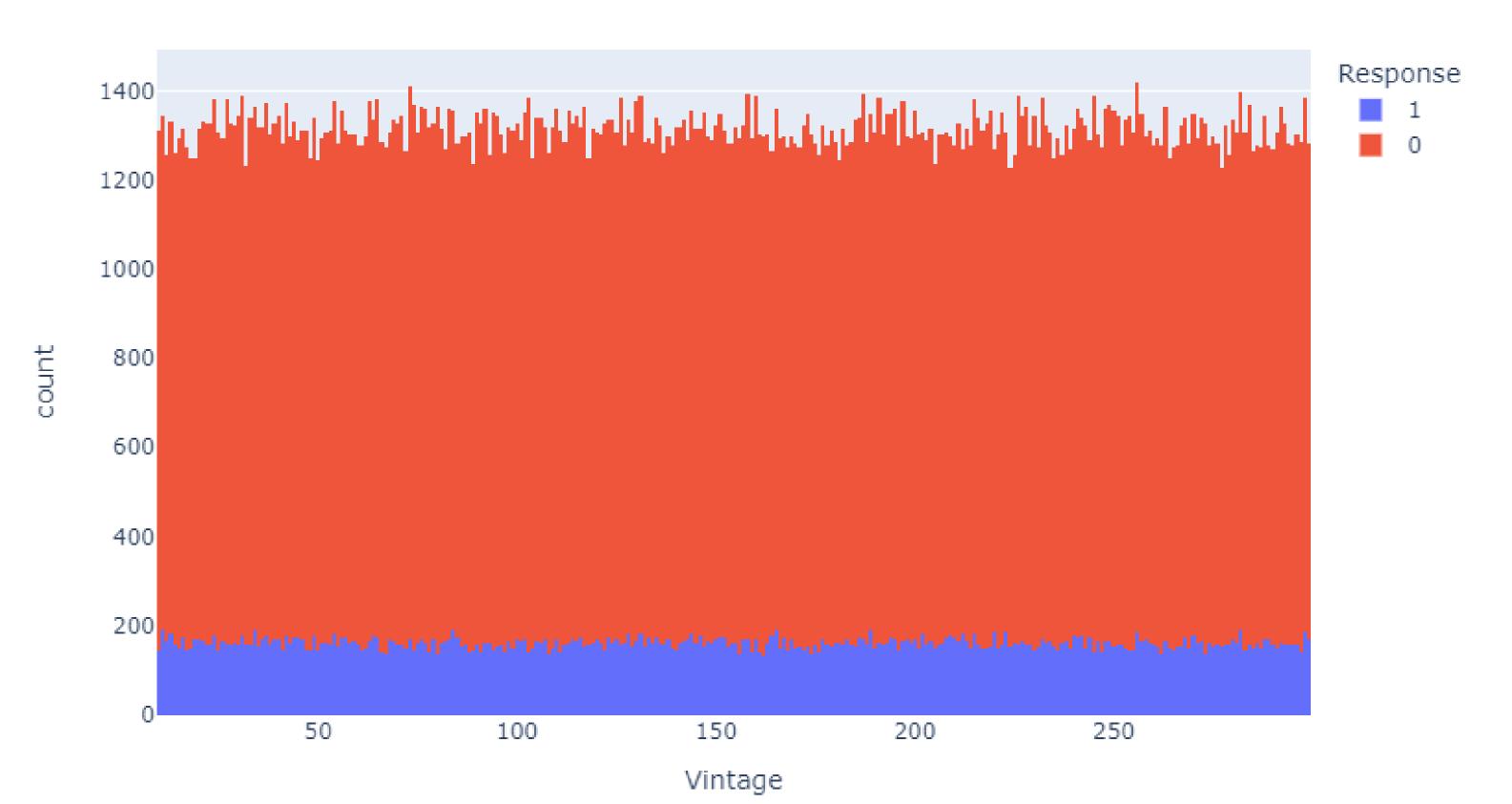
Do different age group have different pattern regarding the decision to buy vehicle insurance?

1 vs 0 Age distribution



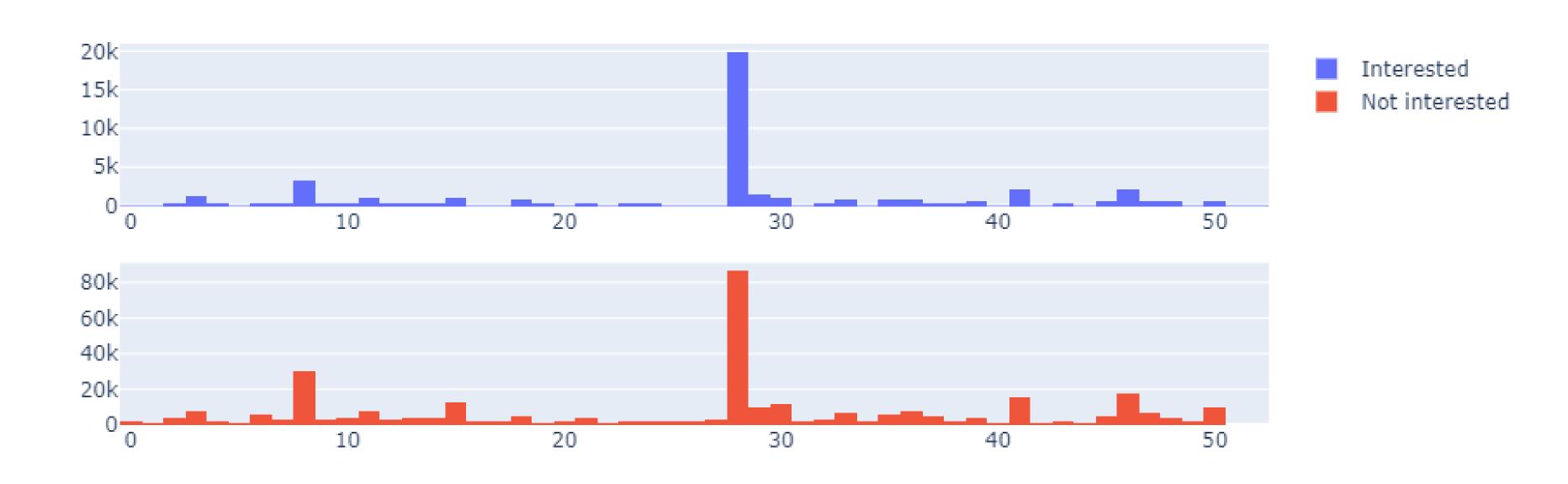
Does longer vintage correlate with higher likelihood to buy vehicle insurance?

Vintage distribution with regard to response



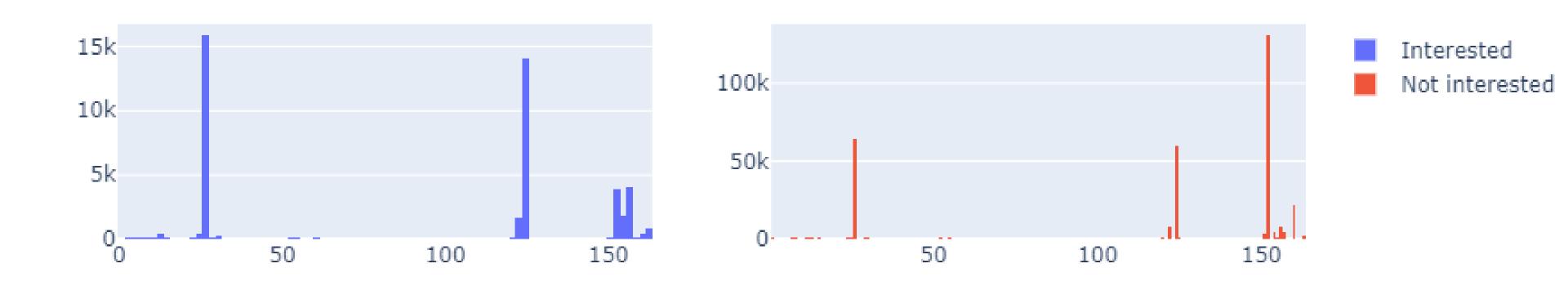
Which region codes would the campaign for increasing sales of vehicle insurance be targeted?

1 vs 0 in Region Code



Which channel would be the optimal one for increasing the sales of vehicle insurance?

1 vs 0 in Policy Sales Channel



Key Findings:

- Traits of potential customers:
- Previously **not granted** insurance policy and had **vehicle damage**
- Male customers in the middle-aged group with vehicle age ranging from 1-2 years
- Customers coming from region code 28

Exploratory Data Analysis

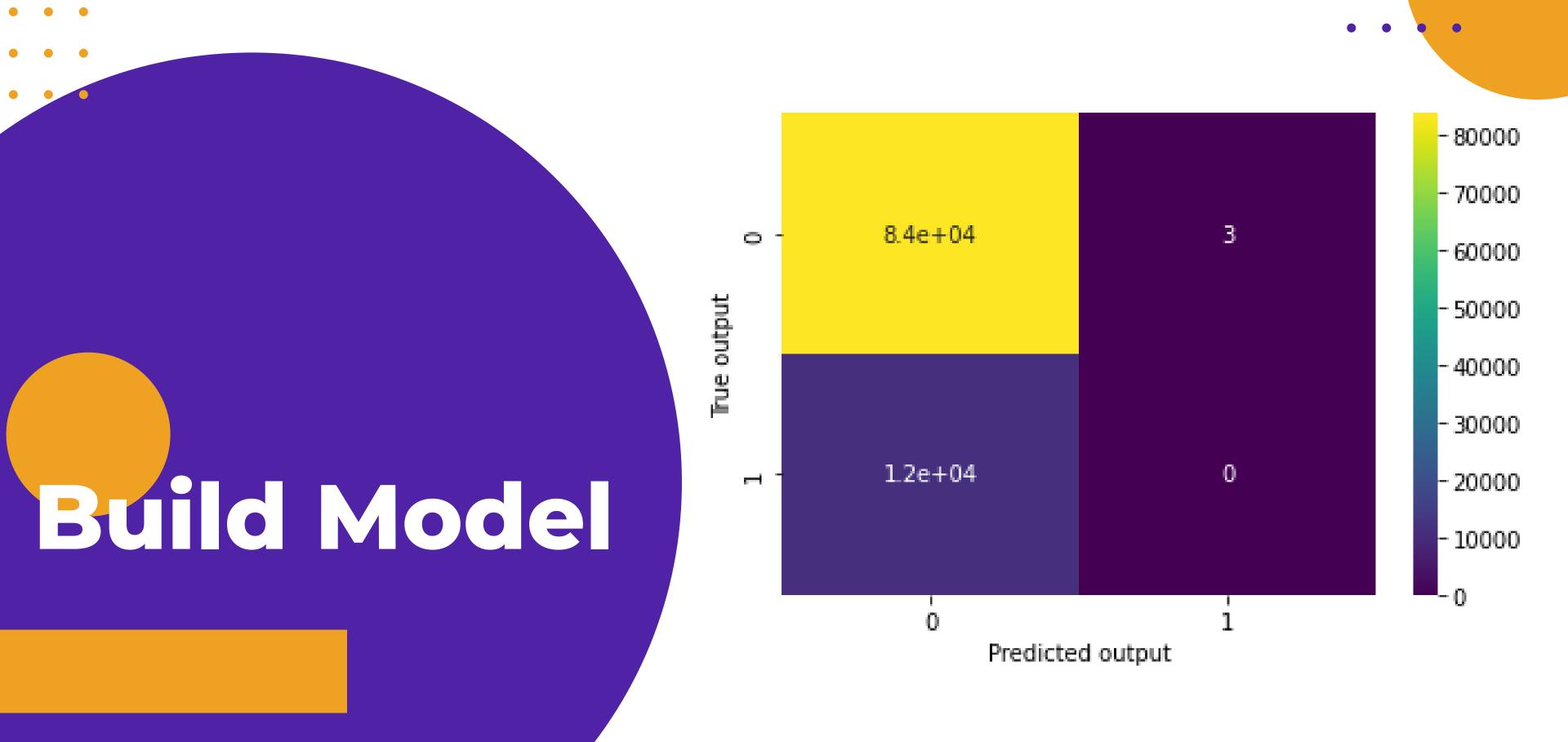
Data Preprocessing

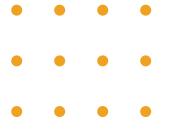
Methods:

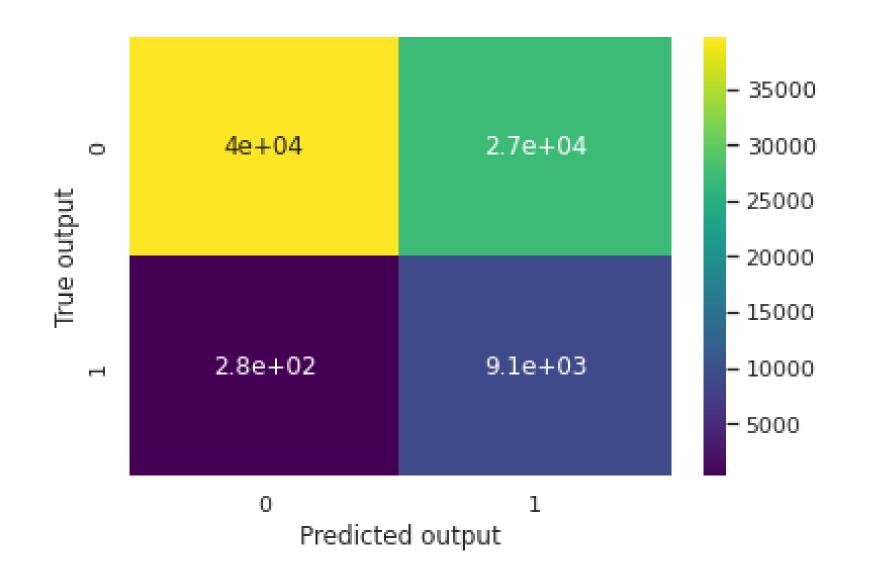
- 1. Convert to numerical value
- region code
- 2. One-hot encode the categorical features
- Vehicle Age, Vehicle_Damage, Gender
- 3. Apply min-max scaler to the continuous values
- Age, vintage, Annual_Premium
- 4. Drop insignificant columns
- id, Policy_Sales_Channel, Driving_License

Gender -	1	0.15	-0.018	0.03	-0.082	0.092	0.0037	-0.11	-0.0025	0.052	0.15	-0.17	0.043
Age -	0.15	1	-0.08	0.3	-0.25	0.27	0.068	-0.58	-0.0013	0.11	0.69	-0.79	0.22
Driving_License -	-0.018	-0.08	1	-0.016	0.015	-0.017	-0.012	0.044	-0.00085	0.01	-0.037	0.04	-0.0062
Region_Code -	0.03	0.3	-0.016	1	-0.18	0.19	0.29	-0.28	0.00037	0.12	0.26	-0.33	0.15
Previously_Insured -	-0.082	-0.25	0.015	-0.18	1	-0.82	0.0043	0.22	0.0025	-0.34	-0.28	0.36	-0.19
Vehicle_Damage -	0.092	0.27	-0.017	0.19	-0.82	1	0.0093	-0.22	-0.0021	0.35	0.28	-0.37	0.21
Annual_Premium -	0.0037	0.068	-0.012	0.29	0.0043	0.0093	1	-0.11	-0.00061	0.023	-0.0025	-0.023	0.062
Policy_Sales_Channel -	-0.11	-0.58	0.044	-0.28	0.22	-0.22	-0.11	1	1.8e-06	-0.14	-0.51	0.57	-0.15
Vintage -	-0.0025	-0.0013	-0.00085	0.00037	0.0025	-0.0021	-0.00061	1.8e-06	1	-0.0011	-0.0026	0.0024	0.0006
Response -	0.052	0.11	0.01	0.12	-0.34	0.35	0.023	-0.14	-0.0011	1	0.16	-0.21	0.11
Vehicle_Age_1-2 Year -	0.15	0.69	-0.037	0.26	-0.28	0.28	-0.0025	-0.51	-0.0026	0.16	1	-0.92	-0.22
Vehicle_Age less than 1 Year -	-0.17	-0.79	0.04	-0.33	0.36	-0.37	-0.023	0.57	0.0024	-0.21	-0.92	1	-0.18
Vehicle_Age over 2 Years -	0.043	0.22	-0.0062	0.15	-0.19	0.21	0.062	-0.15	0.0006	0.11	-0.22	-0.18	1
Feature Selection	Gender -	Age -	Driving_License -	Region_Code -	Previously_Insured -	Vehicle_Damage -	Annual_Premium -	Policy_Sales_Channel -	Vintage –	Response -	Vehicle_Age_1-2 Year -	Vehicle_Age less than 1 Year -	Vehicle_Age over 2 Years -

- 1.00 - 0.75 - 0.50 - 0.25 - 0.00 - -0.25 - -0.50





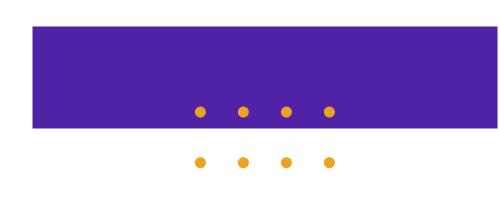


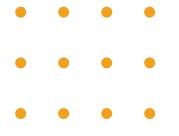
Baseline Model

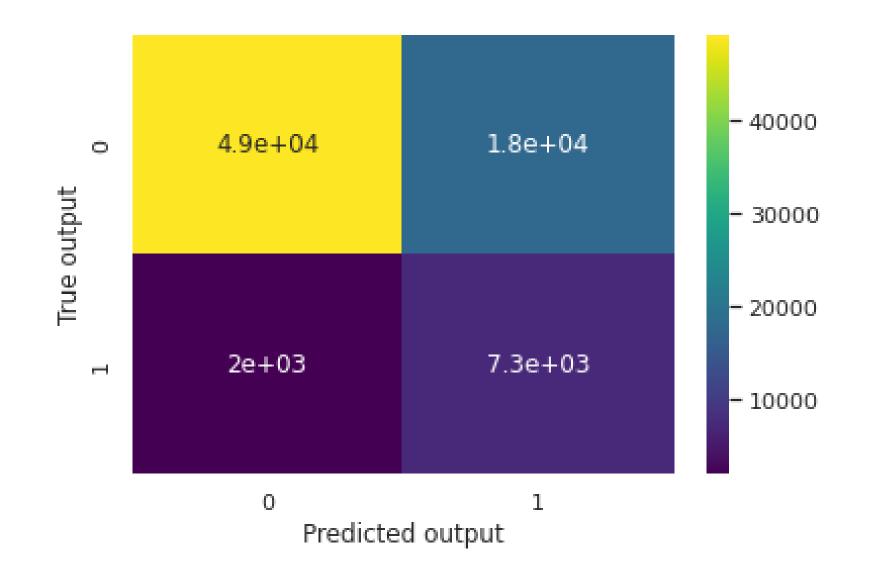
Logistic Regression is chosen as the baseline model

F1 score: 0.397

Recall: 0.97





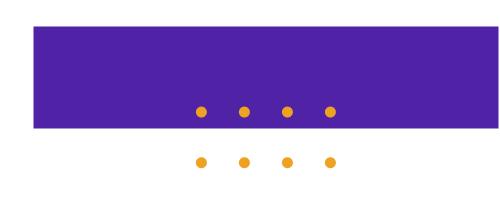


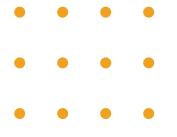
Ensemble Method

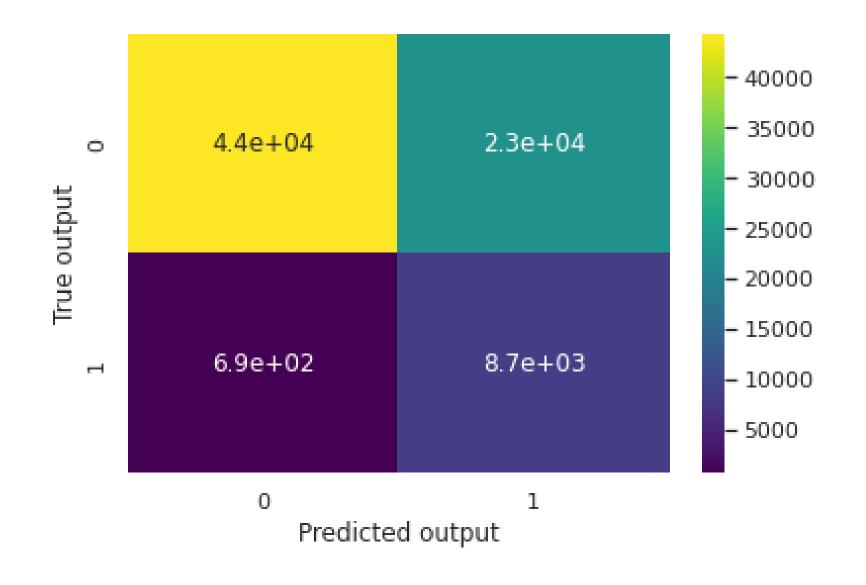
XGBoost Classifier is firstly chosen

F1 score: 0.427

Recall: 0.785





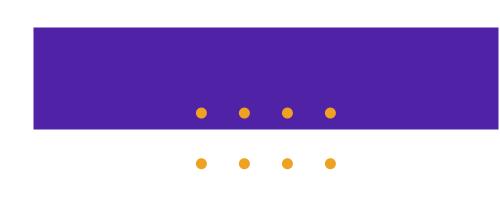


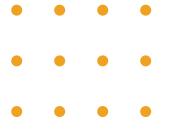
Ensemble Method

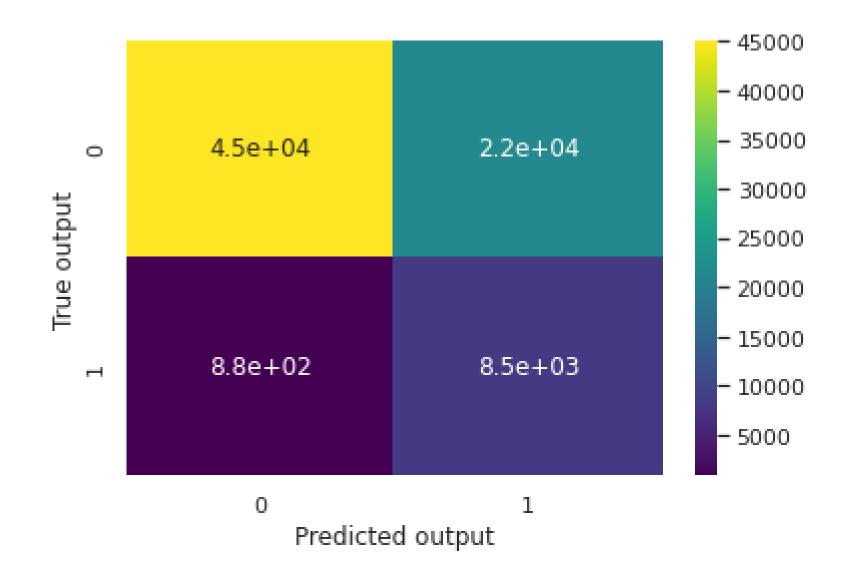
Random Forest Classifier is then chosen

F1 score: 0.427

Recall: 0.926





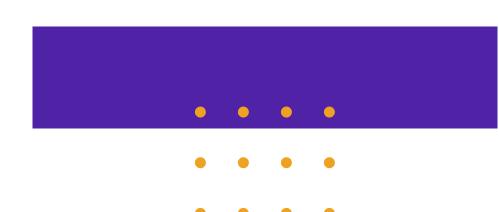


Ensemble Method

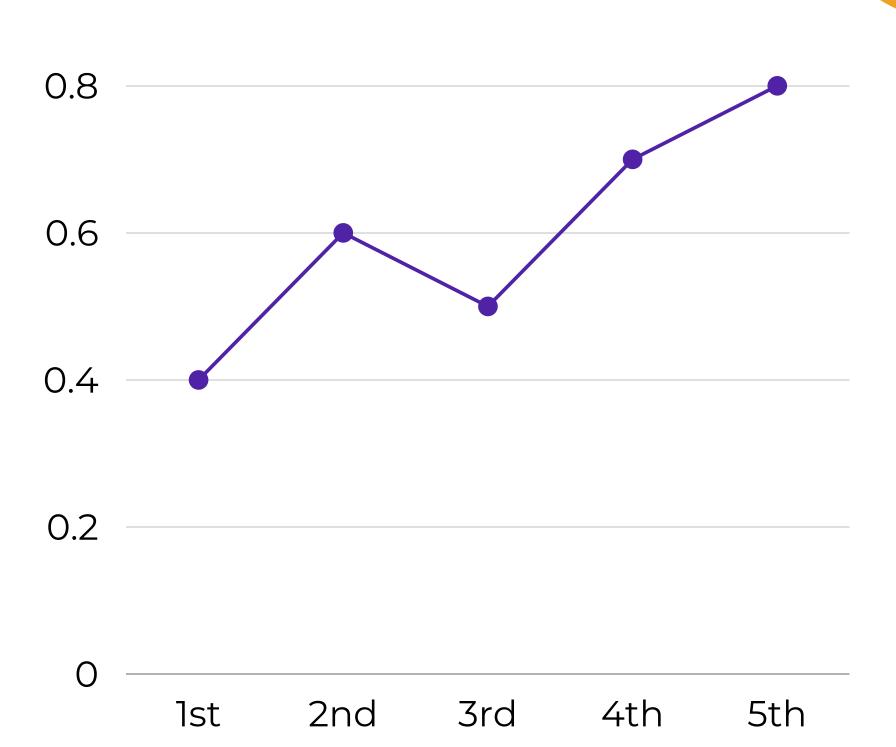
Light Gradient Boosting Classifier is then chosen

F1 score: 0.428

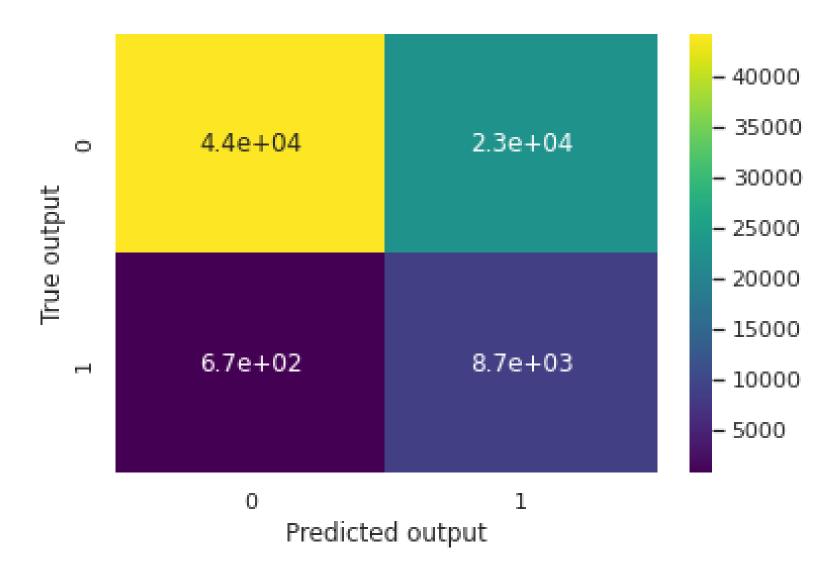
Recall: 0.91



Improvement to the model



Oversampling (RFC)

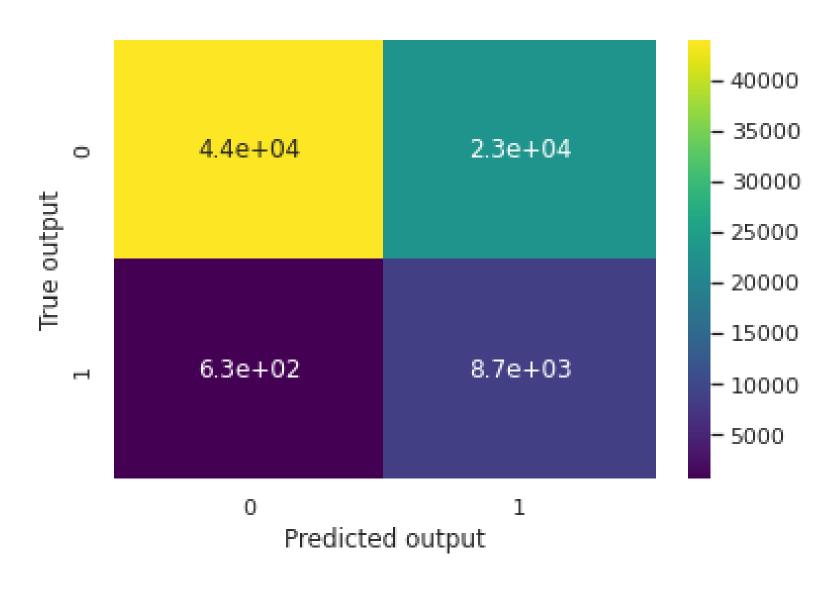


F1 score: 0.426

Recall: 0.93

Precision: 0.276

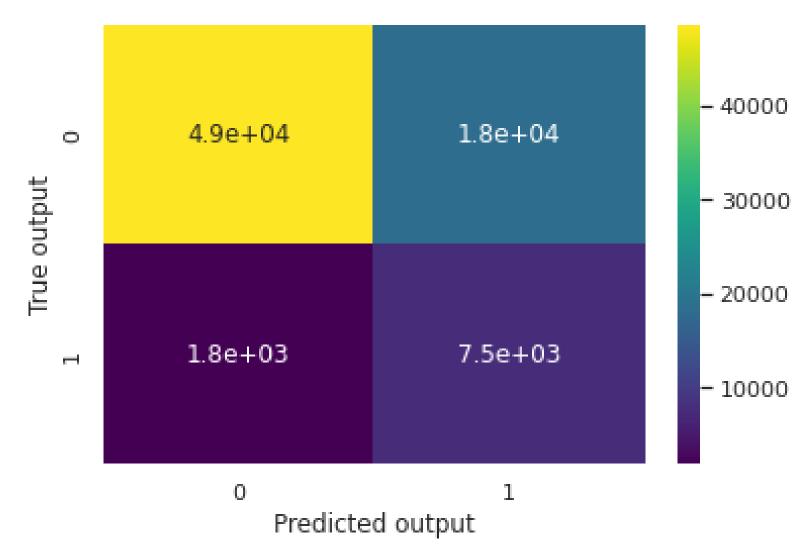
Undersampling (RFC)



F1 score: 0.425

Recall: 0.932

Oversampling (XGBoost)

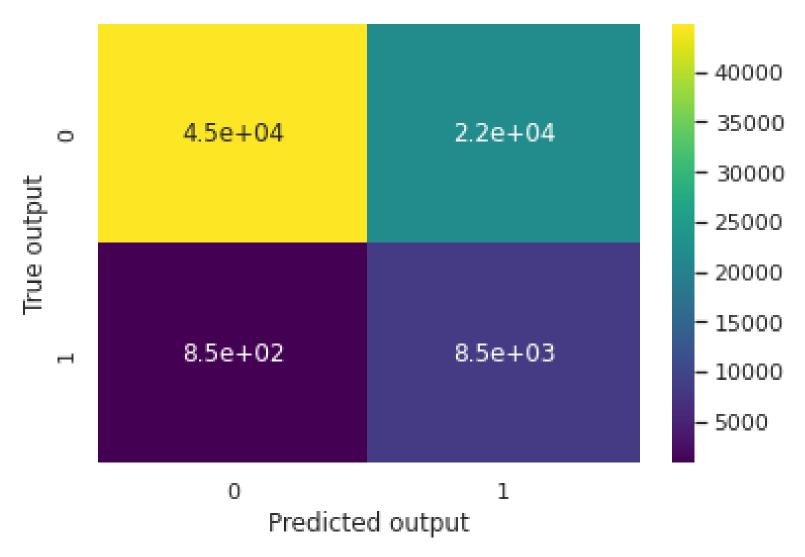


F1 score: 0.428

Recall: 0.907

Precision: 0.29

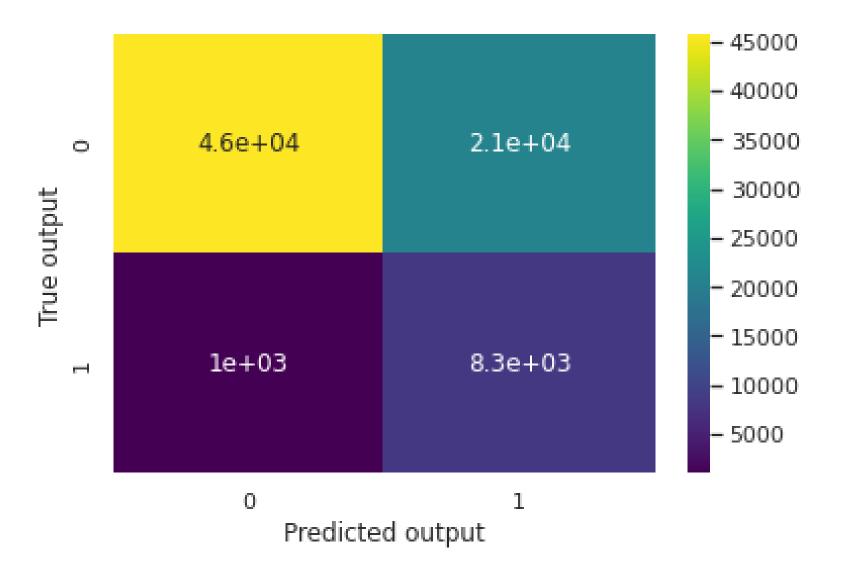
Undersampling (XGBoost)



F1 score: 0.426

Recall: 0.91

Oversampling (LGBM)

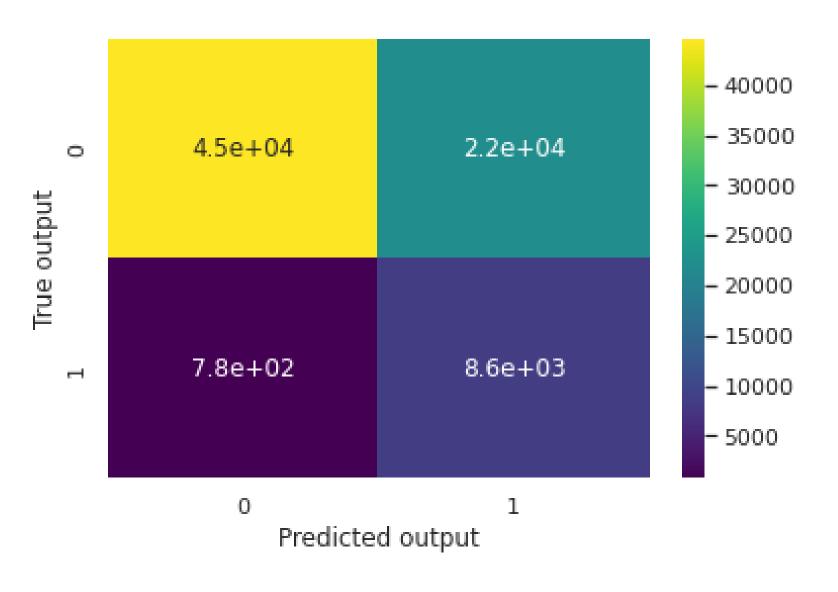


F1 score: 0.43

Recall: 0.892

Precision: 0.283

Undersampling (LGBM)



F1 score: 0.426

Recall: 0.917

