Eunbin Yoo

Data Science Project

Car Insurance Sales Prediction

Data visualizations using Matplotlib and Seaborn and Binary classification model building

Problem Statement

An insurance company seeks to expand its offerings by introducing vehicle insurance to its existing health insurance customers. To achieve this, we aim to build a predictive model to determine whether policyholders from the past year will be interested in purchasing vehicle insurance.

Project Task:

By analyzing a dataset that includes demographics (gender, age, region code type), vehicle details (vehicle age, damage history), and policy specifics (premium amount, sourcing channel), we aim to develop a model that accurately forecasts which customers are likely to purchase vehicle insurance.

Dataset Used

Data Dictionary

Variable	Definition	Key
id		
Gender	Gender of a customer	'Male', 'Female'
Age	Age of a customer	
Driving_License	Does a customer have a driving license?	0=No, 1=Yes
Region_Code		
Previously_Insured	Is a customer insured previously?	0=No, 1=Yes
Vehicle_Age	Age of a customer's vehicle	'1-2 Year', 'Less than 1 Year', 'More than 2 Years'
Annual_Premium	Annual premium ammount	
Policy_Sales_Channel	Sales channel	
Vintage	Days customer associated for with company	
Response		

- The dataset used in this project is “Health Insurance Cross Sell dataset” uploaded on Kaggle under a public domain.
- This dataset involves information about demographics, vehicles and insurance policy. Each row represents a health insurance customer.

Approach

The project involves these steps:

1. **Data Pre-processing**: Data cleaning, handling missing values, feature scaling, and encoding categorical variables
2. **Exploratory Data Analysis** (EDA): Descriptive analysis and visualizations
3. **Feature Engineering**: Creation, transformation and selection of features
4. **Model Building**: Regression algorithms for binary classification model, such as logistic linear regression and random forests classifier
5. **Model Evaluation**: Area Under The Curve (AUC) Score

Key Findings 1

The cross-sell analysis provided valuable insights into the dynamics of customer behavior and preferences.

1. The findings highlight notable demographic trends, such as the higher interest among male customers compared to females and the age distribution of interested customers, particularly the significant proportion within the 40-49 age group. Understanding these demographic nuances is essential for refining marketing strategies and optimizing outreach efforts.

2. The observed negative correlation between age and policy sales channel preference offers valuable guidance for optimizing sales channel allocation and customer outreach strategies. By aligning outreach channels with customer preferences, insurers can enhance engagement and conversion rates.

Key Findings 2

The cross-sell analysis provided valuable insights into the dynamics of customer behavior and preferences.

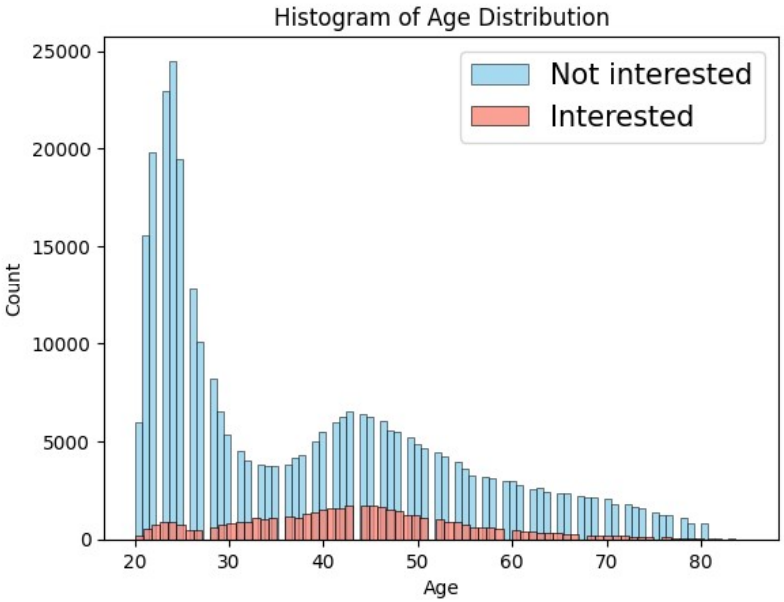
3. The utilization of a Light Gradient Boosting Model has significantly enhanced our predictive accuracy, with an impressive AUC score of 86%, indicating its robust performance in identifying potential customers interested in vehicle insurance.

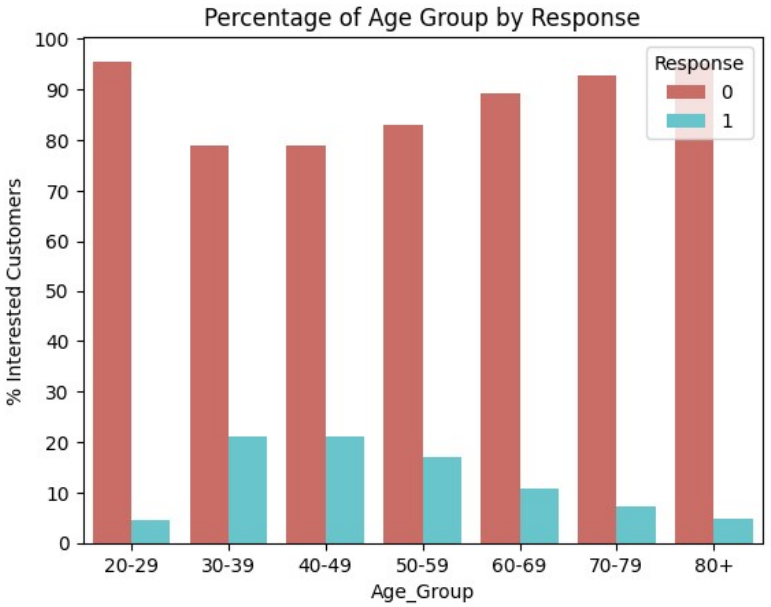
4. The analysis underscores the significance of annual premium as the most influential determinant in predicting customer interest in additional insurance products. This insight suggests the importance of tailored pricing strategies and personalized offerings to effectively target and engage potential customers.

Visualizations

Data visualizations using
matplotlib and seaborn

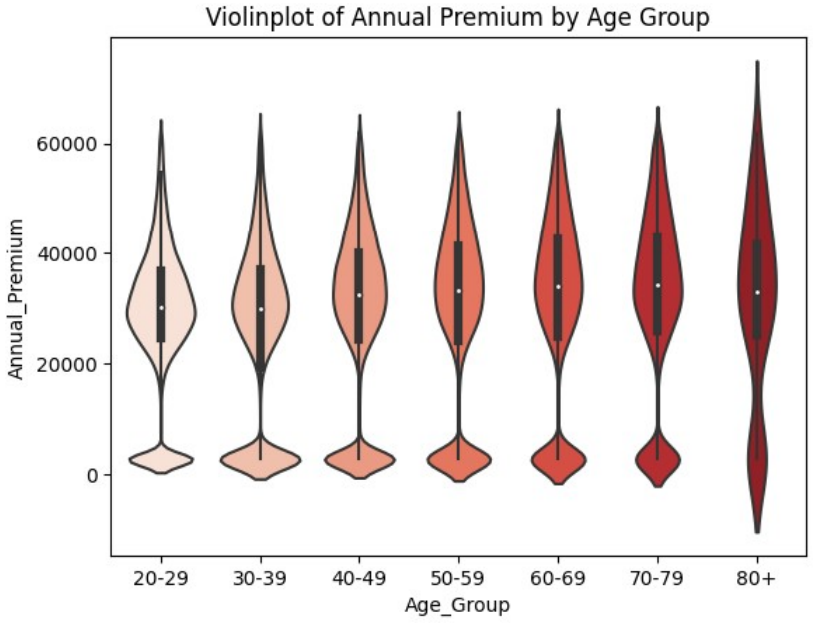
Age Distribution of Customers





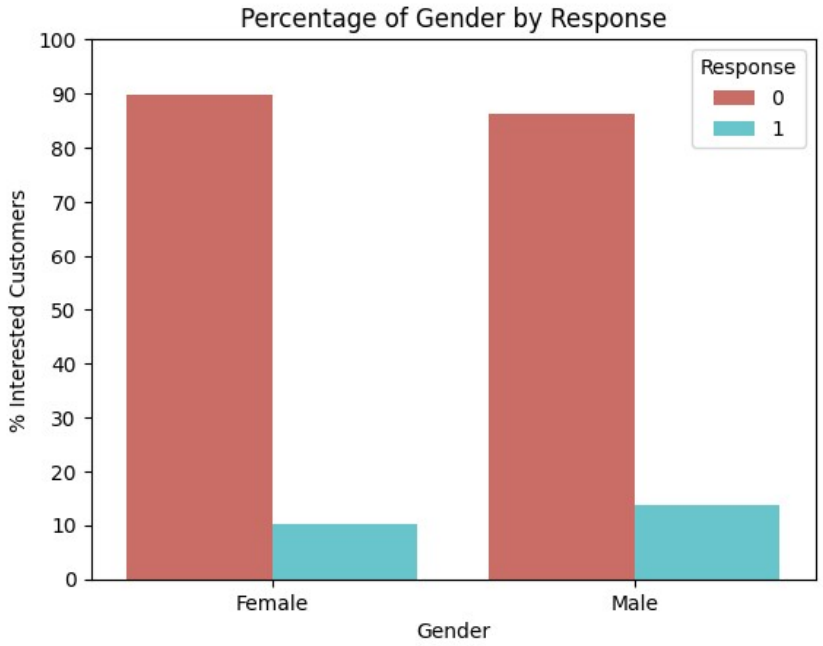
The bar chart shows that 40-49 is the age group with the largest proportion of interested customers (20.1%).
The proportion of the customers interested in vehicle insurance is smaller in older age groups.

Age Vs. Annual Premium Relationship



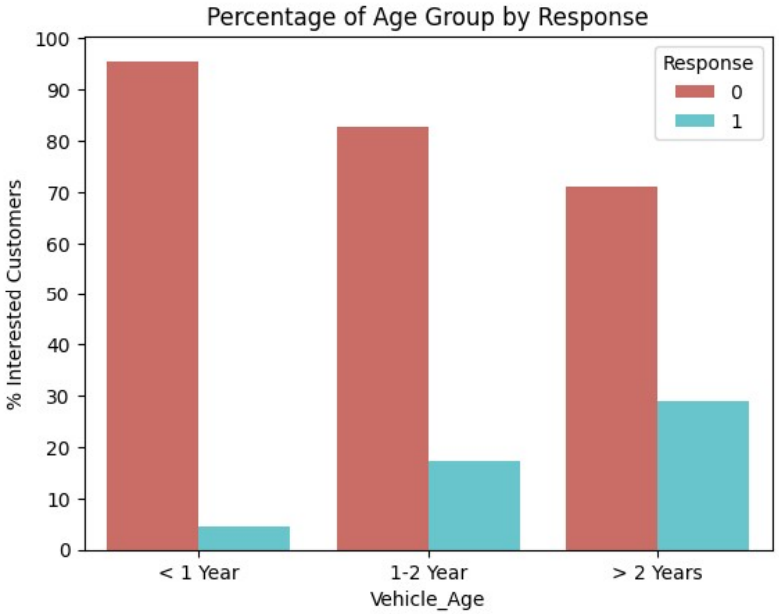
The violin plot shows that older customers pay higher annual premium than younger customers on average.

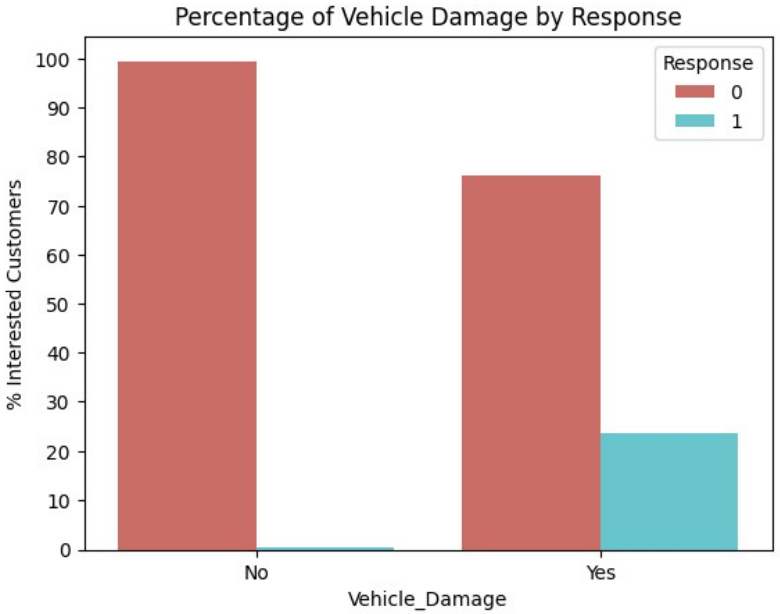
Gender Distribution



About 3% more male customers are interested in vehicle insurance than female customers.

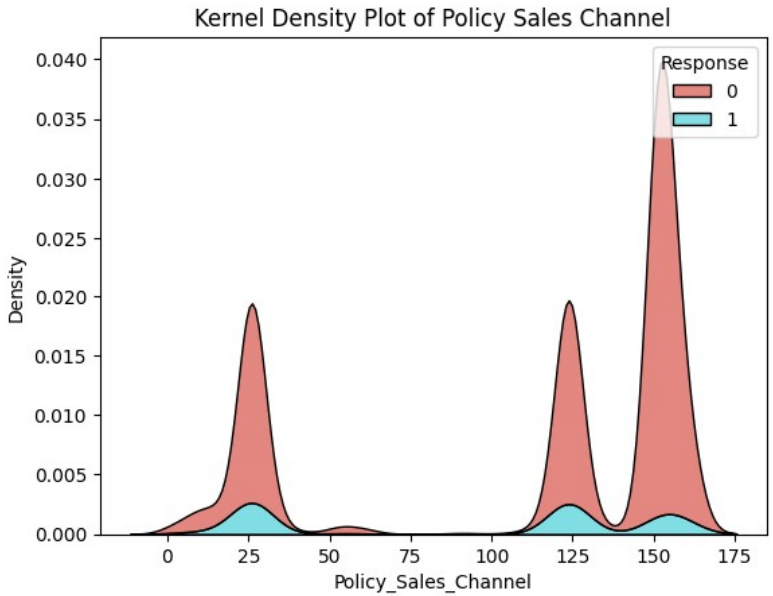
Vehicle Damage and Age

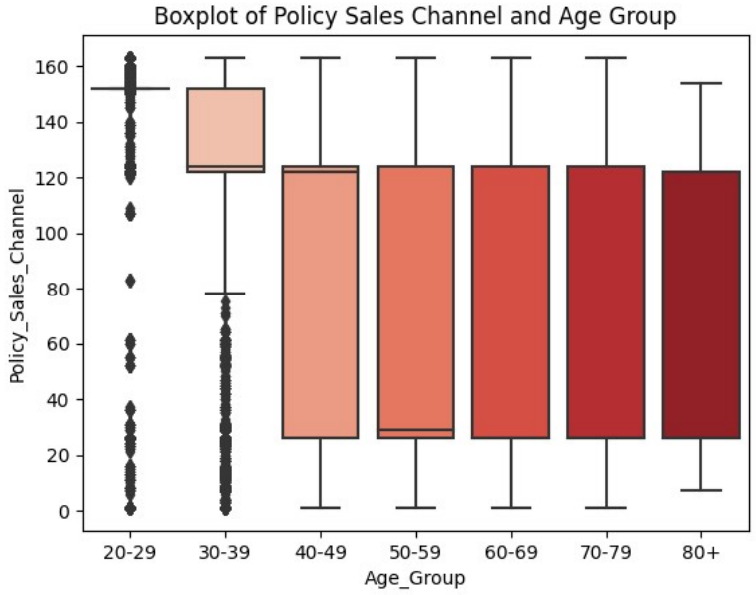




Customers whose vehicles are older or have damage are more interested in vehicle insurance. Only few customers who does not have damage on their vehicles are interested.

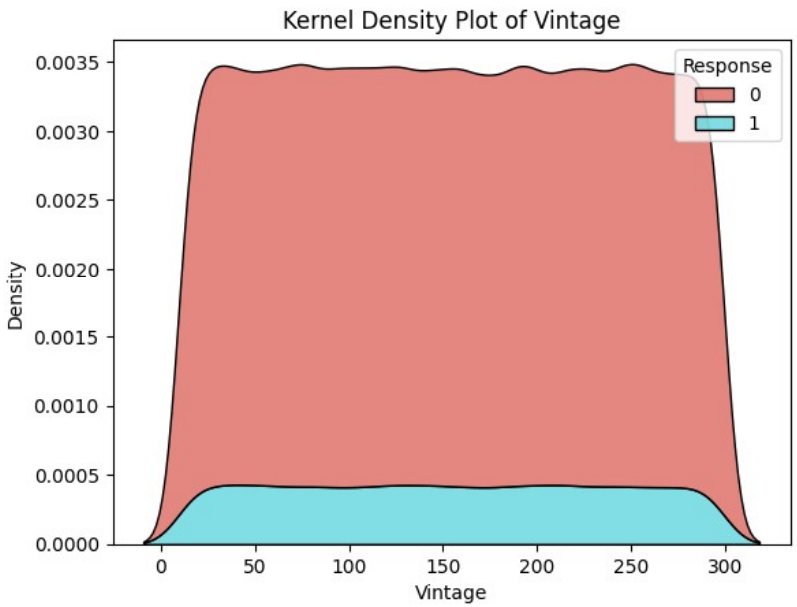
Policy Sales Channel Distribution

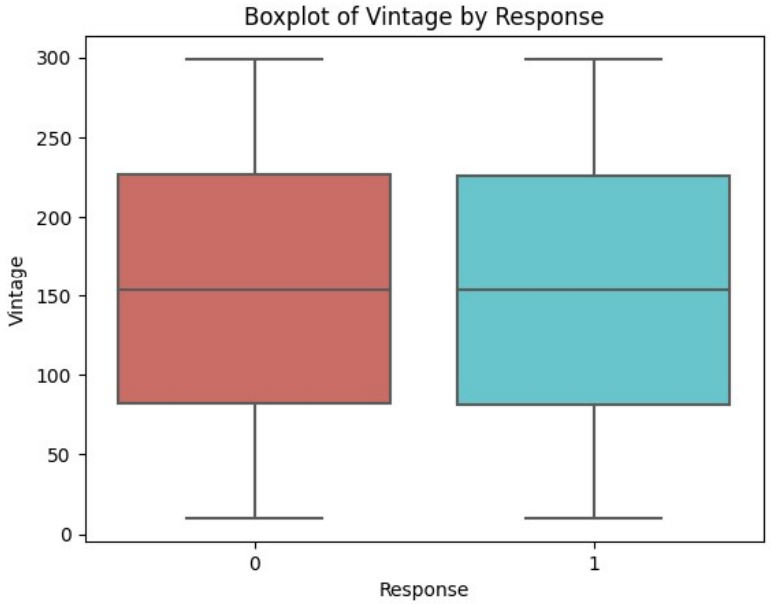




The plots suggest that there is a difference in preference for Policy Sales Channel between different age groups. Younger age groups tend to prefer higher Policy Sales Channel code, while older age groups tend to prefer Lower Policy Channel code.

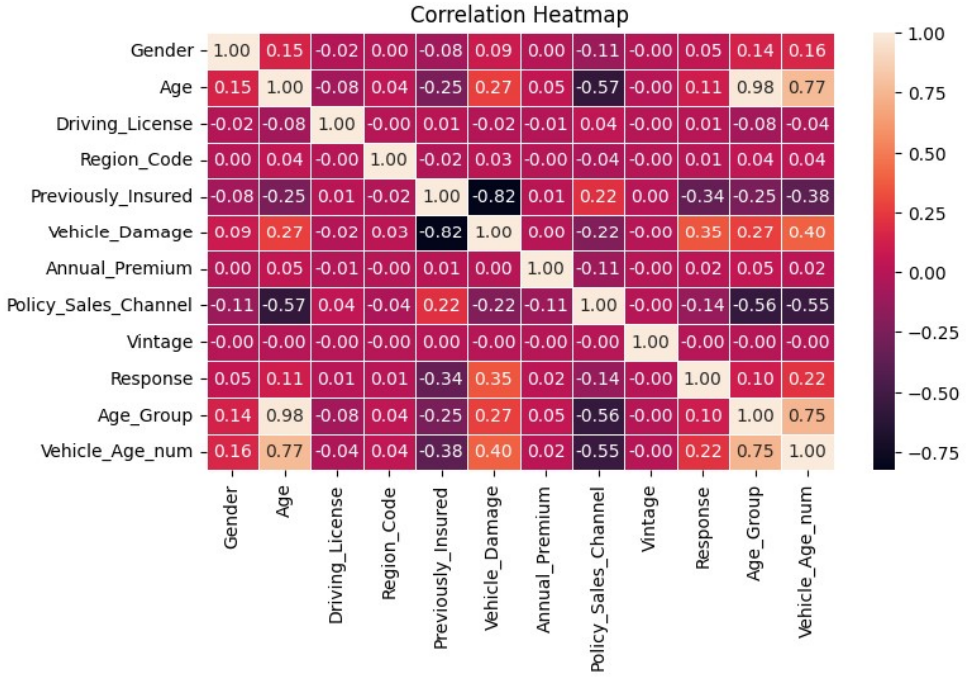
Vintage Distribution





The density plot shows that there is almost no difference in the average number of days customers have been associated with the company between those who are interested in Vehicle insurance and those who are not.

Correlation Analysis

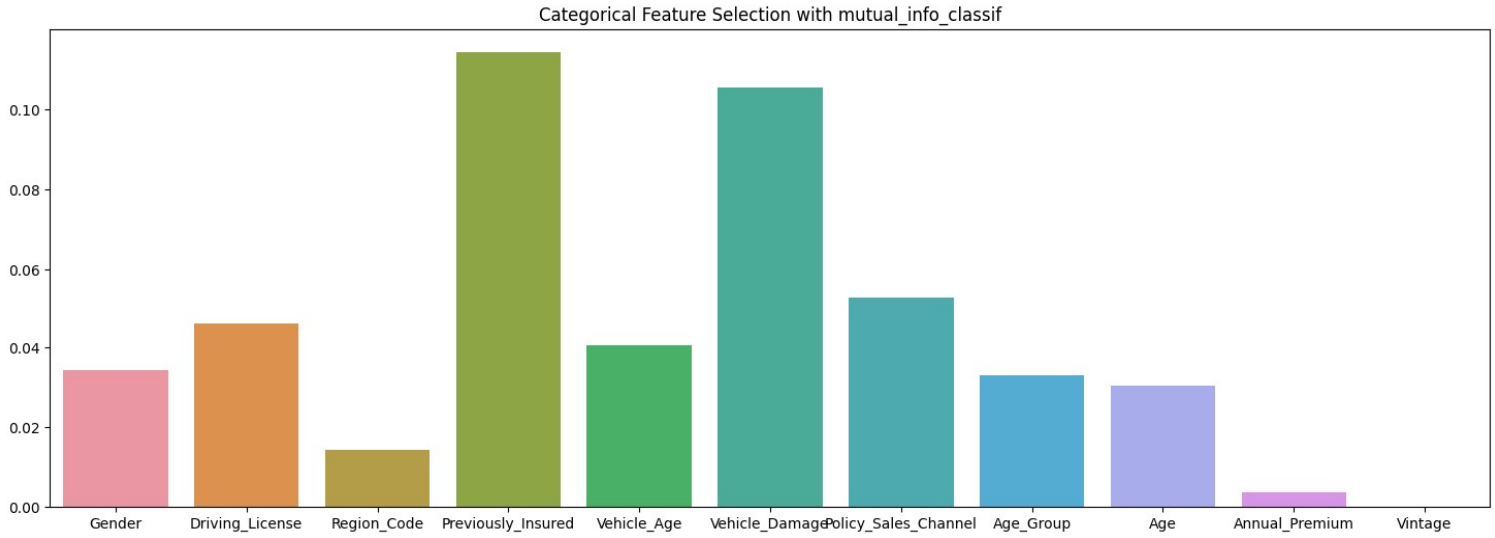


- The heat map shows a moderate negative correlation (-0.57) between Age and Policy Sales Channel, suggesting that older age group prefers to be outreached by lower Policy Sales Channel code.
- There is a fair positive correlation (0.75) between Vehicle Age and Age. Older age group has older vehicles.

Model Building and Evaluation

Binary classification model using various machine learning algorithms

Feature Selection



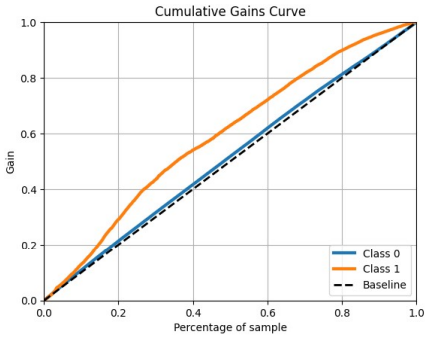
According to K-scores, ‘Age’ and ‘Vintage’ are dropped from the final dataset.

- The chart shows that adding age groups as a new feature brings small improvement.
- Vintage has the lowest k-score.

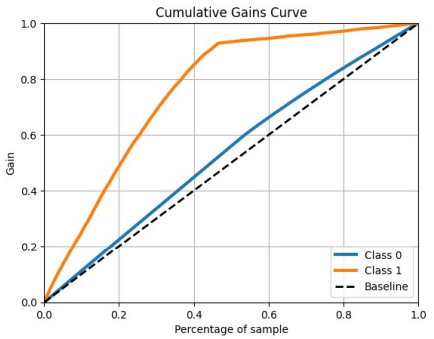
Cumulative Gains Curve

Logistic Regression, Random Forest Classifier, XGBoost Classifier and LGBM Model are built.

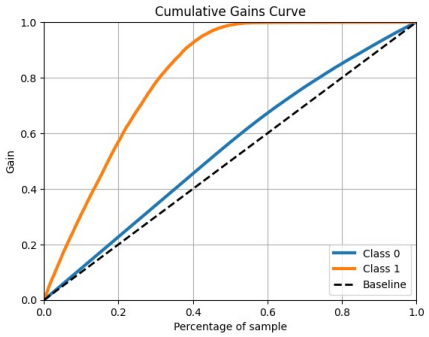
Model 1:
Logistic Regression



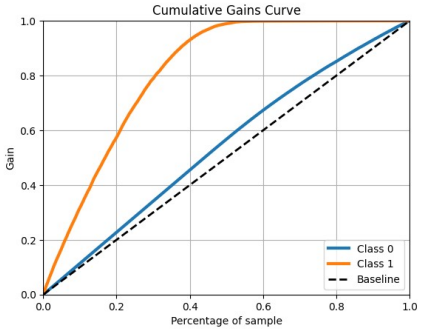
Model 2:
Random Forest Classifier



Model 3:
XGBoost Classifier



Model 4:
LGBM



Model Performance

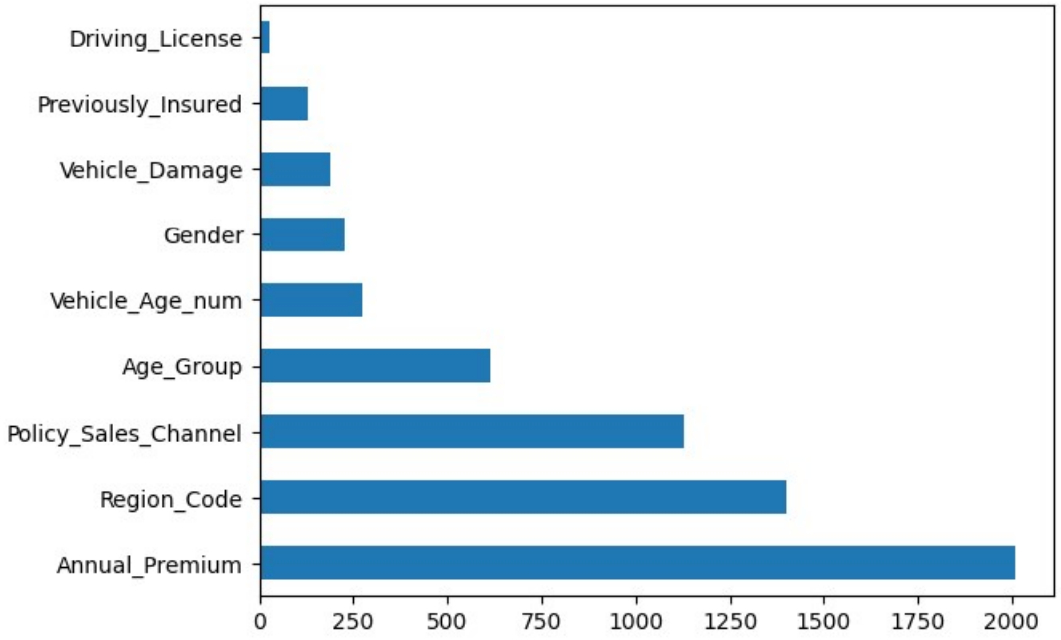
I chose the model with the best cost-benefit ratio (higher score, lower size, higher speed).

Model Name	Accuracy Balanced	Precision @K Mean	Recall @K Mean	ROC AUC Score	Top K Score
LGBMClassifier	0.504510	0.280886	0.926421	0.855200	0.877840
XGBClassifier	0.517045	0.279206	0.920880	0.850569	0.875968
RandomForestClassifier	0.582536	0.259607	0.856238	0.800011	0.844800
LogisticRegression	0.515601	0.240328	0.792660	0.771975	0.872559

Final Model

- The utilization of a Light Gradient Boosting Model has significantly enhanced our predictive accuracy, with an impressive AUC score of 86%, indicating its robust performance in identifying potential customers interested in vehicle insurance.
- Final model is a Light Gradient Boosting Model with the highest AUC score of 0.86.

Feature Importance



- The model suggests that Annual Premium is the most important feature in predicting whether a customer is interested in Vehicle insurance.
- Driving License is the least important feature.

Conclusion

Limitations:

The analysis reveals a weakness in the form of data imbalance, where only 12% of customers are interested in vehicle insurance. Addressing this imbalance through oversampling techniques may resolve this issue and potentially improve the overall model performance by providing more balanced representation of both interested and non-interested customers.

Conclusions:

In conclusion, the analysis not only provides actionable insights for optimizing cross-selling strategies but also demonstrates the efficacy of advanced modeling techniques in extracting meaningful patterns from complex datasets. These insights can empower insurers to tailor their marketing approaches, deepen customer relationships, and drive business growth in an increasingly competitive marketplace.

Thank you

- Link to Original Dataset: https://www.kaggle.com/datasets/anmolkumar/health-insurance-cross-sell-prediction/data?select=train.csv
- Github Link for code, resources and more information: https://github.com/bebe5004/Eunbin-Yoo-s-Portfolio/tree/main/Car%20Insurance%20Sales%20Prediction
- Github page for more projects: https://github.com/bebe5004/Eunbin-Yoo-s-Portfolio/tree/main?tab=readme-ov-file#readme