Batch details	PGP-DSE-FT-BLR-JAN-2024
Team members	Sudarshan Venkatesh Akhil T Nandini MK Alishanandini Basa Vishnu EM Aditya Singh
Domain of Project	Retail and Marketing
Proposed project title	Car Loan Claim Prediction
Group Number	2
Team Leader	Sudarshan Venkatesh
Mentor Name	Mr. Jatinder Bedi

Abstract

This project aims to develop a robust predictive model for car loan claim likelihood using machine learning techniques. Utilizing a dataset of 10,000 policyholder records, we explored various demographic, vehicle-related, and driving history features to identify key factors influencing claim probability. The project involved extensive data preprocessing, exploratory data analysis, and the implementation of multiple classification algorithms.

Our methodology included handling missing values, feature engineering, and addressing class imbalance. We performed in-depth exploratory data analysis to uncover patterns and relationships within the data. Several machine learning models were developed and compared, including Logistic Regression, Random Forest, Gradient Boosting, Decision Trees, K-Nearest Neighbors, XGBoost, and AdaBoost.

The best-performing model was the Gradient Boosting Classifier, achieving an accuracy of 85.30% on the test set. Key predictors of loan claims included credit score, annual mileage, and driving history. The project demonstrated the potential for machine learning to significantly enhance risk assessment in the loan industry.

This report details our approach, findings, and recommendations for implementing predictive modeling in car loan claim assessment. The insights gained from this project have the potential to optimize pricing strategies, improve risk assessment, and streamline underwriting processes for loan companies.

1. Introduction

1.1 Background

The car loan industry relies heavily on accurate risk assessment to maintain profitability and offer fair premiums to policyholders. Traditionally, actuarial methods have been used to evaluate risk, but these approaches can be enhanced by leveraging machine learning techniques to analyze complex patterns in policyholder data.

In recent years, the availability of large datasets and advancements in machine learning algorithms have opened new possibilities for predictive modeling in the loan sector. By accurately predicting the likelihood of a claim, loan companies can:

- Optimize pricing strategies
- Improve risk assessment
- Enhance customer segmentation
- Streamline the underwriting process

This project focuses on developing a predictive model for car loan claims using a dataset of 10,000 policyholders, aiming to revolutionize how loan companies approach risk management and policy pricing in the automotive sector.

1.2 Objectives

The primary objectives of this project are:

1. To develop a machine learning model that accurately predicts the likelihood of car loan claims.
2. To identify the most important factors influencing loan claim probability.
3. To compare the performance of various machine learning algorithms in the context of loan claim prediction.
4. To provide actionable insights for loan companies to improve their risk assessment and pricing strategies.
5. To demonstrate the potential of data-driven decision-making in the loan industry.

1.3 Scope

This project encompasses the following scope:

- Analysis of a dataset containing 10,000 policyholder records with 18 features.
- Comprehensive data preprocessing, including handling missing values and feature engineering.
- Exploratory data analysis to uncover patterns and relationships within the data.
- Implementation and comparison of multiple machine learning algorithms for claim prediction.
- Evaluation of model performance using various metrics such as accuracy, precision, recall, and F1-score.
- Interpretation of results and provision of business insights for the loan industry.

The project does not include real-time implementation or integration with existing loan systems, focusing instead on the development and evaluation of predictive models.

2. Dataset Description

2.1 Data Source

The dataset used in this project was obtained from Kaggle, a popular platform for data science and machine learning datasets. The original owner of the data is Sagnik Roy, and the dataset is titled "Car Loan Data".
Source: https://www.kaggle.com/datasets/sagnik1511/car-loan-data

2.2 Context and Significance

This dataset provides a comprehensive view of car loan policyholders, including various attributes that could potentially influence the likelihood of filing an loan claim. The data is significant for several reasons:

1. It allows for the exploration of relationships between policyholder characteristics and claim likelihood.
2. It provides an opportunity to develop predictive models that can assist loan companies in risk assessment.
3. The dataset includes a mix of demographic, vehicle-related, and driving history features, offering a holistic view of factors influencing loan claims.

2.3 Data Characteristics and Handling

- **Size**: The dataset contains 10,000 policyholder records.
- **Features**: There are 18 predictor variables and 1 target variable.
- **Data types**: The dataset includes a mix of categorical and numerical variables.
- **Missing values**: Some columns (CREDIT_SCORE and ANNUAL_MILEAGE) contain missing values.
- **Class imbalance**: The target variable (OUTCOME) is imbalanced, with a minority class representing claim occurrences.

2.4 Variables and Descriptions

1. AGE: Age group of the policyholder
2. GENDER: Gender of the policyholder
3. RACE: Race of the policyholder
4. DRIVING_EXPERIENCE: Driving experience in years
5. EDUCATION: Education level of the policyholder
6. INCOME: Income category of the policyholder
7. CREDIT_SCORE: Credit score of the policyholder
8. VEHICLE_OWNERSHIP: Whether the policyholder owns the vehicle (1) or not (0)
9. VEHICLE_YEAR: Year of vehicle manufacture
10. MARRIED: Marital status of the policyholder

11. CHILDREN: Number of children
12. POSTAL_CODE: Postal code of the policyholder's residence
13. ANNUAL_MILEAGE: Annual mileage driven
14. VEHICLE_TYPE: Type of vehicle
15. SPEEDING_VIOLATIONS: Number of speeding violations
16. DUIS: Number of DUIs (Driving Under Influence)
17. PAST_ACCIDENTS: Number of past accidents
18. OUTCOME: Whether a claim was filed (1) or not (0)

3. Methodology

3.1 Data Preprocessing

3.1.1 Handling Missing Values

Two features had missing values:

- CREDIT_SCORE: 982 missing values (9.82%)
- ANNUAL_MILEAGE: 957 missing values (9.57%)

To address these missing values, we employed the following strategies:

1. CREDIT_SCORE: Missing values were imputed using the mean credit score grouped by EDUCATION and INCOME.
2. python

```
df['CREDIT_SCORE'] =
df['CREDIT_SCORE'].fillna(df.groupby(['EDUCATION','INCOME'])['CREDIT_SCORE'].transform('mean')
```

3. ANNUAL_MILEAGE: Missing values were imputed using the mean annual mileage grouped by VEHICLE_TYPE and VEHICLE_YEAR.
4. python

```
df['ANNUAL_MILEAGE'] =
df['ANNUAL_MILEAGE'].fillna(df.groupby(['VEHICLE_TYPE','VEHICLE_YEAR'])['ANNUAL_MILEAGE'].transform('mean'))
```

3.1.2 Handling Outliers

Outliers were identified and treated for numerical features using the Interquartile Range (IQR) method:

python

```
for i in list(df.select_dtypes(np.number).columns):
    q1 = df[i].quantile(0.25)
    q3 = df[i].quantile(0.75)
    iqr = q3 - q1
    ul = q3 + (1.5 * iqr)
    ll = q1 - (1.5 * iqr)
    df[i] = df[i].apply(lambda x: ll if x < ll else ul if x > ul else x)
```

3.1.3 Feature Engineering

Several new features were created to capture additional information:

1. RISK_SCORE: Calculated as the sum of SPEEDING_VIOLATIONS, DUIS, and PAST_ACCIDENTS.

```
df['RISK_SCORE'] = df['SPEEDING_VIOLATIONS'] + df['DUIS'] + df['PAST_ACCIDENTS']
```

2. LOCATION: Mapped from POSTAL_CODE to provide more meaningful geographical information.

```
             def map_location(x):
    if x == 10238:
        return 'New York'
    if x == 32765:
        return 'Florida'
    if x == 92101:
        return 'California'
    if x == 21217:
        return 'Mary Land'
df['LOCATION'] = df['POSTAL_CODE'].apply(map_location)
```

3. risk_score_bin: Binned version of RISK_SCORE for easier interpretation.

```
def bin_risk_score(x):
    if x <= 4:
        return 'low risk'
    else:
        return 'high_risk'
df['risk_score_bin'] = df['RISK_SCORE'].apply(bin_risk_score)
```

3.1.4 Encoding Categorical Variables

Categorical variables were encoded using various techniques:

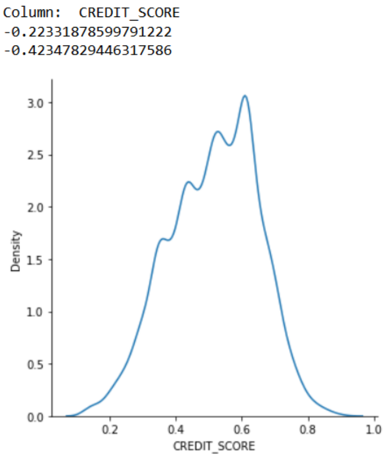
1. One-hot encoding for variables like AGE, GENDER, LOCATION, and DRIVING_EXPERIENCE.
2. Ordinal encoding for EDUCATION and INCOME.
3. Binary encoding for VEHICLE_YEAR and VEHICLE_TYPE.

3.2 Exploratory Data Analysis

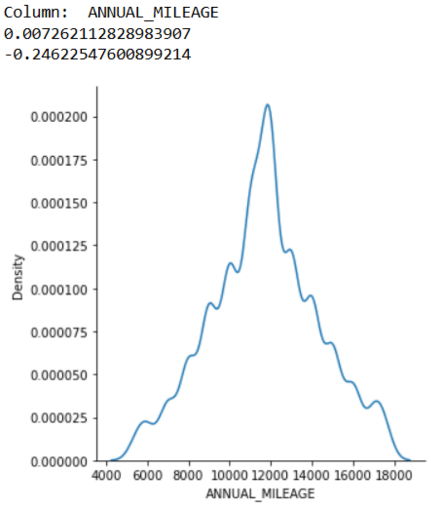
3.2.1 Univariate Analysis

Numerical Columns Analyzed:

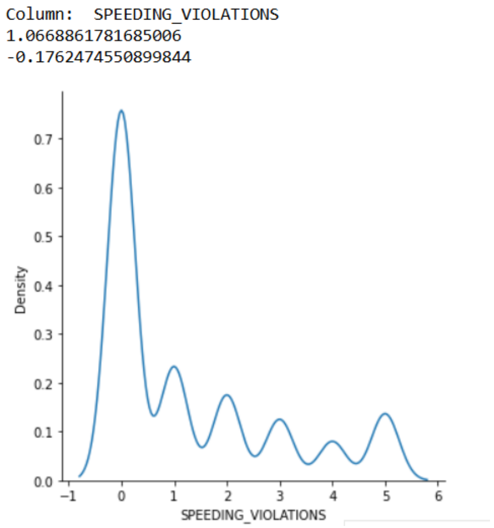
- CREDIT_SCORE
- VEHICLE_OWNERSHIP
- MARRIED
- CHILDREN
- ANNUAL_MILEAGE
- SPEEDING_VIOLATIONS
- DUIS
- PAST_ACCIDENTS
- OUTCOME



```
Column:  CREDIT_SCORE
-0.22331878599791222
-0.42347829446317586
```



```
Column:  ANNUAL_MILEAGE
0.007262112828983907
-0.24622547600899214
```



```
Column:  SPEEDING_VIOLATIONS
1.0668861781685006
-0.1762474550899844
```

For each numerical column, we examined:

- Distribution using histograms and kernel density estimation plots
- Descriptive statistics (mean, median, standard deviation, etc.)
- Skewness and kurtosis

3.2.2 Summary of Findings

Skewness:

- CREDIT_SCORE: Slightly negatively skewed (-0.22)
- VEHICLE_OWNERSHIP: Negatively skewed (-0.86)
- ANNUAL_MILEAGE: Approximately symmetric (0.007)
- SPEEDING_VIOLATIONS: Positively skewed (1.07)
- PAST_ACCIDENTS: Positively skewed (1.46)
- OUTCOME: Positively skewed (0.81)

Outliers:

- ANNUAL_MILEAGE: 3% outliers
- SPEEDING_VIOLATIONS: 6% outliers
- DUIS: 19% outliers
- PAST_ACCIDENTS: 3% outliers

Data Concentration:

- CREDIT_SCORE: Most values concentrated between 0.4 and 0.6
- ANNUAL_MILEAGE: Majority of values between 10,000 and 15,000 miles

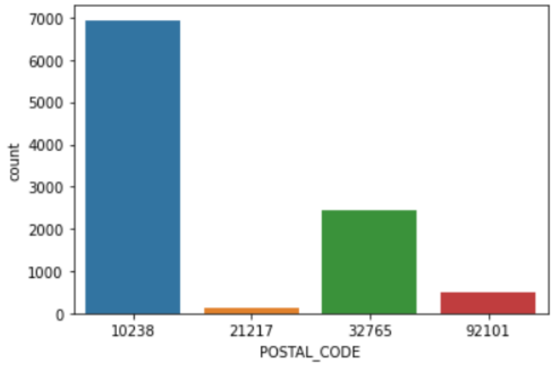
Insights by Feature:

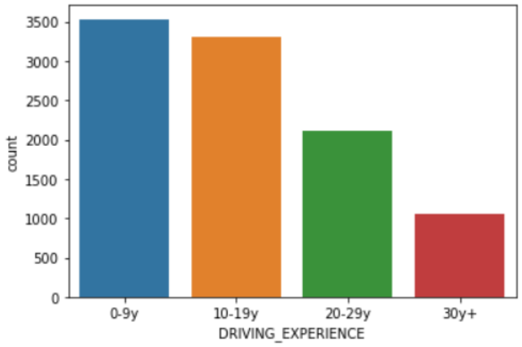
- CREDIT_SCORE: Relatively normal distribution with a slight negative skew
- VEHICLE_OWNERSHIP: Binary feature with more people owning vehicles
- MARRIED: Binary feature with a fairly even split

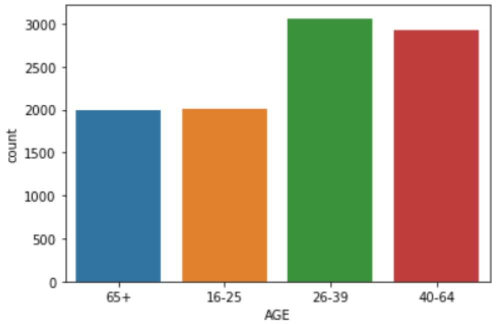
- CHILDREN: Right-skewed distribution, many policyholders with 0 or 1 child
- ANNUAL_MILEAGE: Approximately normal distribution
- SPEEDING_VIOLATIONS: Right-skewed, most policyholders have 0 or few violations
- DUIS: Highly right-skewed, majority have 0 DUIs
- PAST_ACCIDENTS: Right-skewed, most policyholders have 0 or few past accidents
- OUTCOME: Imbalanced binary feature, more non-claim (0) than claim (1) instances

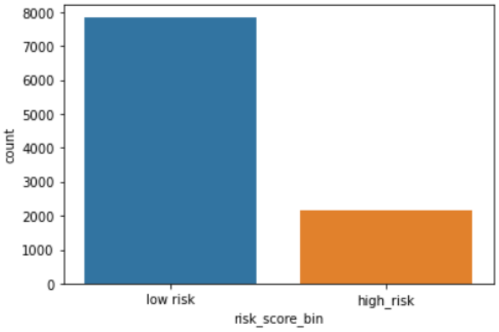
3.2.3 Categorical Columns Analyzed:

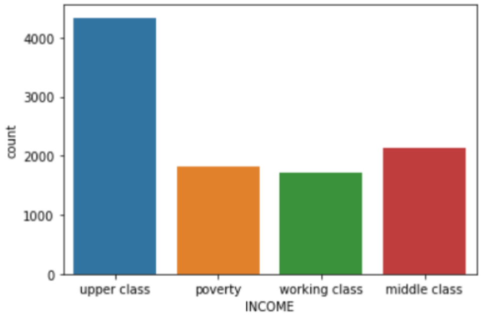
- AGE
- GENDER
- RACE
- DRIVING_EXPERIENCE
- EDUCATION
- INCOME
- VEHICLE_YEAR
- VEHICLE_TYPE
- LOCATION

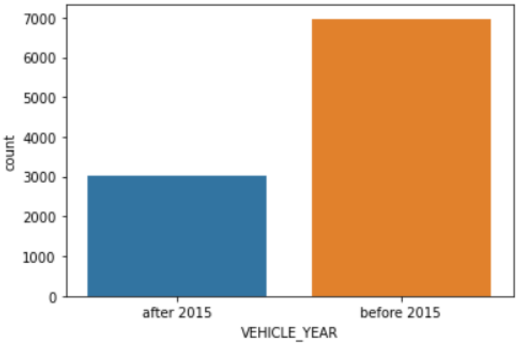












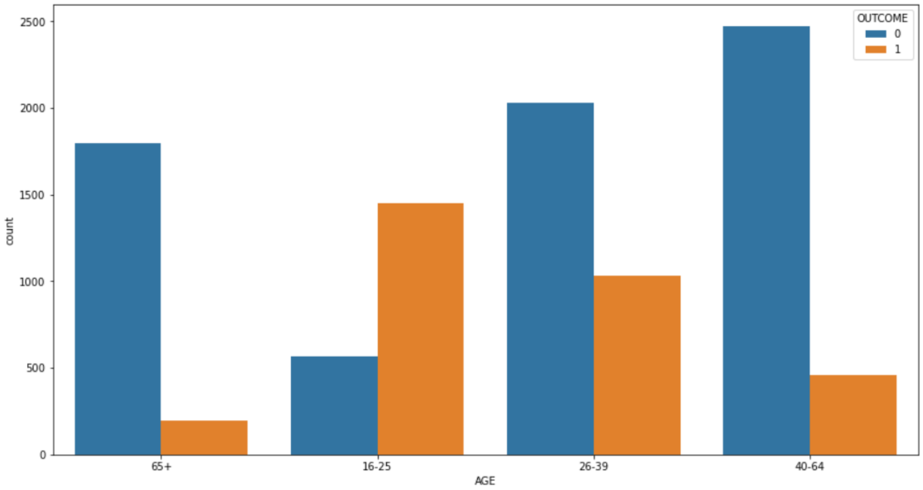
Insights:

- AGE: Most policyholders in the 26-39 and 40-64 age groups
- GENDER: Fairly balanced distribution
- RACE: Majority category is the most common
- DRIVING_EXPERIENCE: More policyholders with 0-9 years of experience
- EDUCATION: Various levels represented, with high school and university being common
- INCOME: Working class and middle class are the most represented categories
- VEHICLE_YEAR: More vehicles from before 2015
- VEHICLE_TYPE: Sedan is the most common vehicle type
- LOCATION: New York has the highest representation, followed by Florida

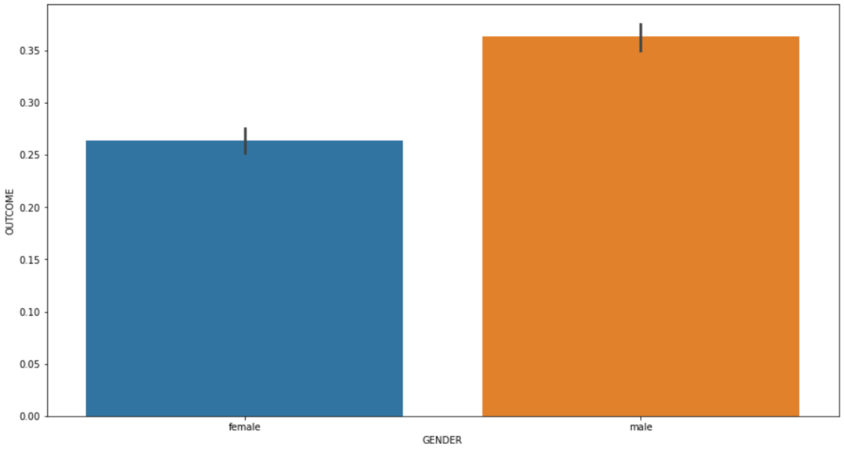
3.2.4 Bivariate Analysis

We conducted several bivariate analyses to understand the relationships between variables:

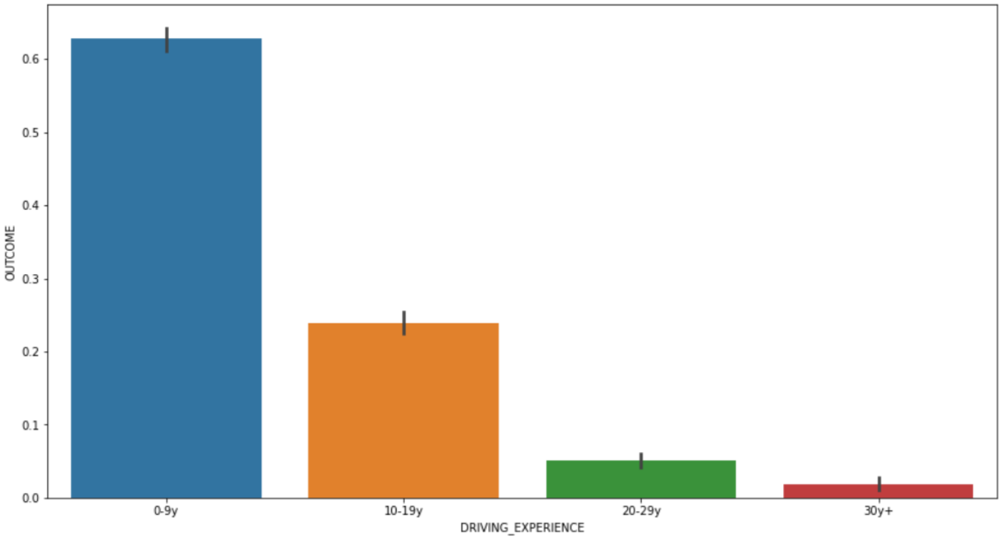
1. AGE vs. OUTCOME:
    - Middle-aged groups (26-39 and 40-64) showed a higher proportion of claims
    - Younger (16-25) and older (65+) groups had lower claim rates



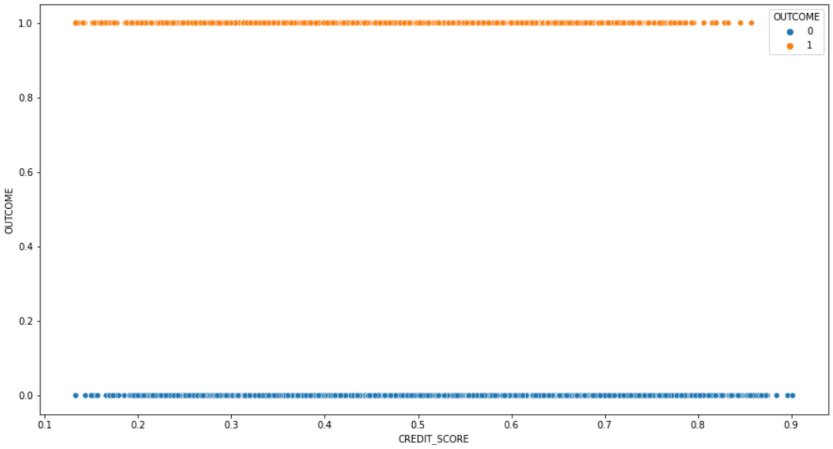
2. GENDER vs. OUTCOME:
    - Slight difference in claim rates between genders
    - Males showed a marginally higher claim rate



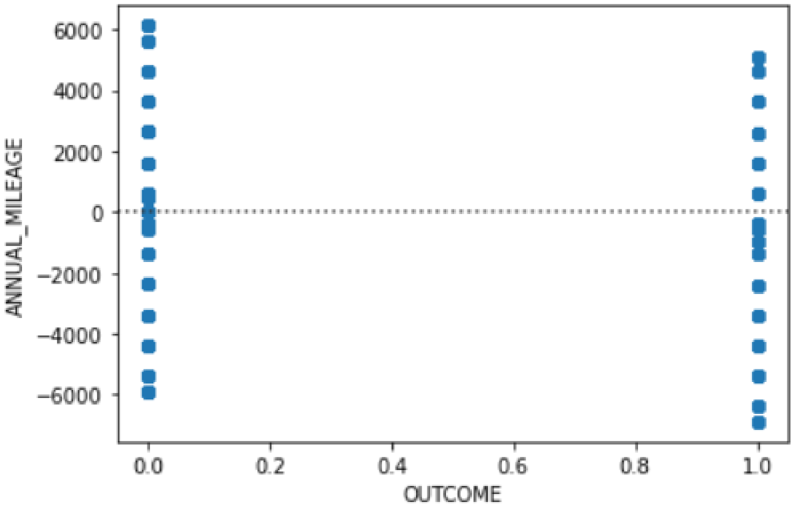
3. DRIVING_EXPERIENCE vs. OUTCOME:
    - Inverse relationship observed
    - Policyholders with less driving experience had higher claim rates



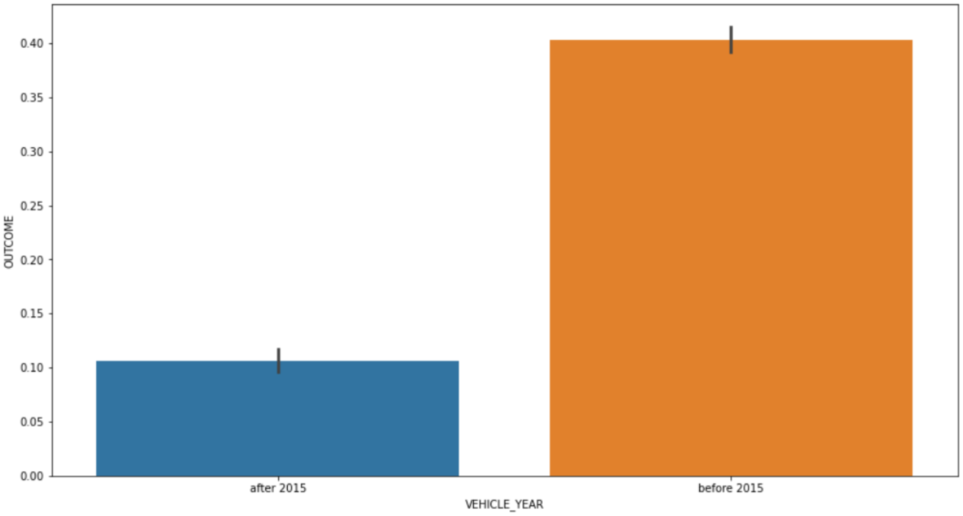
4. CREDIT_SCORE vs. OUTCOME:
    - Negative correlation observed
    - Lower credit scores associated with higher claim probabilities



5. ANNUAL_MILEAGE vs. OUTCOME:
    - Positive correlation
    - Higher annual mileage associated with increased claim likelihood



6. VEHICLE_TYPE vs. OUTCOME:
    - Sports cars showed higher claim rates compared to sedans

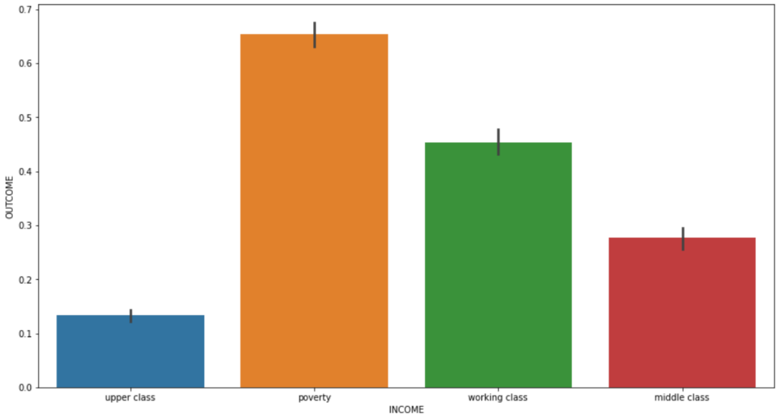


7. SPEEDING_VIOLATIONS vs. OUTCOME:

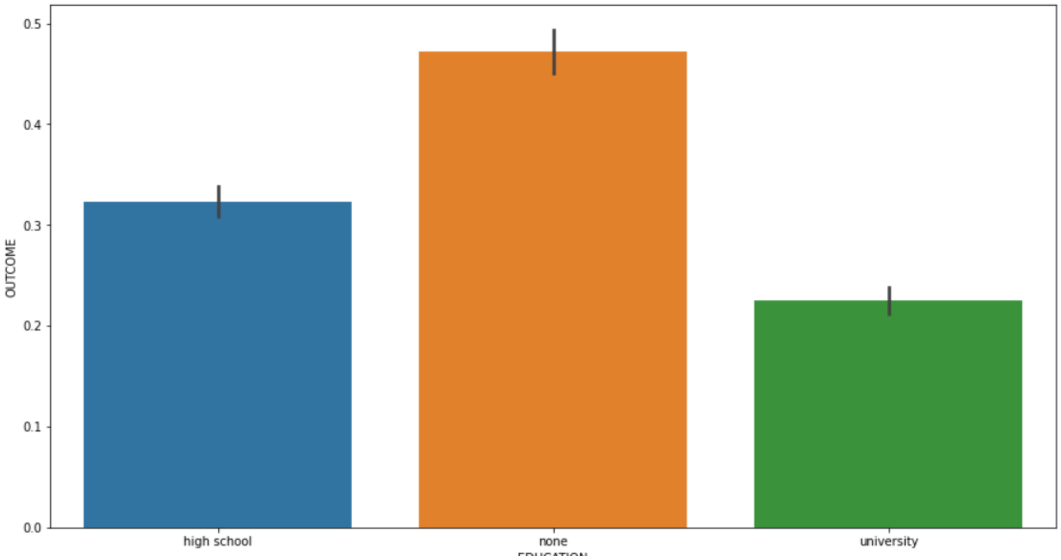
- Positive correlation
- More speeding violations associated with higher claim probabilities

8. PAST_ACCIDENTS vs. OUTCOME:
    - Strong positive correlation
    - More past accidents strongly associated with higher claim rates

9. INCOME vs. OUTCOME:
    - Lower income categories showed slightly higher claim rates



10. EDUCATION vs. OUTCOME:
    - Lower education levels associated with marginally higher claim rates



These bivariate analyses provided valuable insights into the factors influencing loan claim likelihood, guiding feature selection and model development in subsequent stages.

4. Model Building and Evaluation

4.1 Model Selection

Based on the nature of our problem (binary classification) and the characteristics of our dataset, we selected the following machine learning algorithms for evaluation:

1. Logistic Regression
2. Random Forest Classifier
3. Gradient Boosting Classifier
4. Decision Tree Classifier
5. K-Nearest Neighbors Classifier
6. XGBoost Classifier
7. AdaBoost Classifier

These algorithms were chosen for their effectiveness in handling binary classification problems and their ability to capture different types of relationships in the data.

4.2 Model Training

The dataset was split into training (70%) and testing (30%) sets:

```
x = df1.drop('OUTCOME', axis=1)
y = df1['OUTCOME']
xtrain, xtest, ytrain, ytest = train_test_split(x, y, train_size=0.7, random_state=100)
```

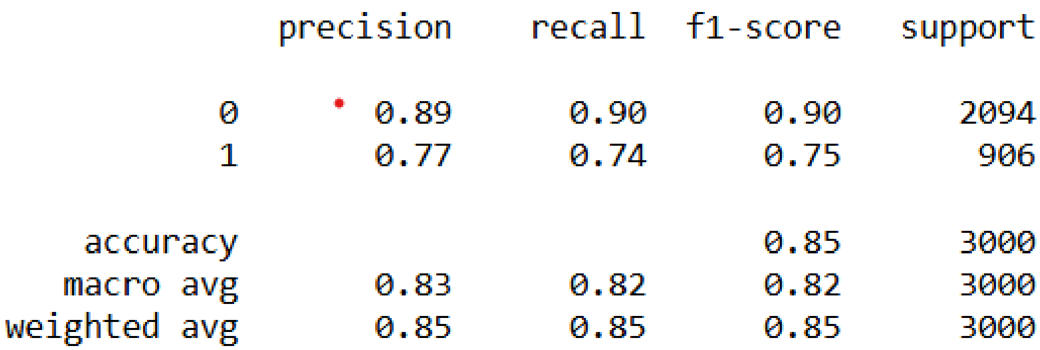
Each model was trained on the training set using default parameters initially. For selected models, hyperparameter tuning was performed using GridSearchCV to optimize performance.

4.3 Model Building_ Base Model

As it is a classification model, we have considered Logistic Regression as our base model

```
from sklearn.linear_model import LogisticRegression

LR2=LogisticRegression()

lr_model=LR2.fit(xtrain,ytrain)

ypred=lr_model.predict(xtest)

ytrain_pred=lr_model.predict(xtrain)
```

Classification Report for the above model is :



```
              precision    recall  f1-score   support

           0       0.89      0.90      0.90      2094
           1       0.77      0.74      0.75       906

    accuracy                           0.85      3000
   macro avg       0.83      0.82      0.82      3000
weighted avg       0.85      0.85      0.85      3000
```

4.3 Model Evaluation

Models were evaluated using the following metrics:

1. Accuracy: Overall correctness of the model
2. Precision: Proportion of true positive predictions
3. Recall: Proportion of actual positives correctly identified
4. F1-score: Harmonic mean of precision and recall
5. ROC-AUC: Area under the Receiver Operating Characteristic curve

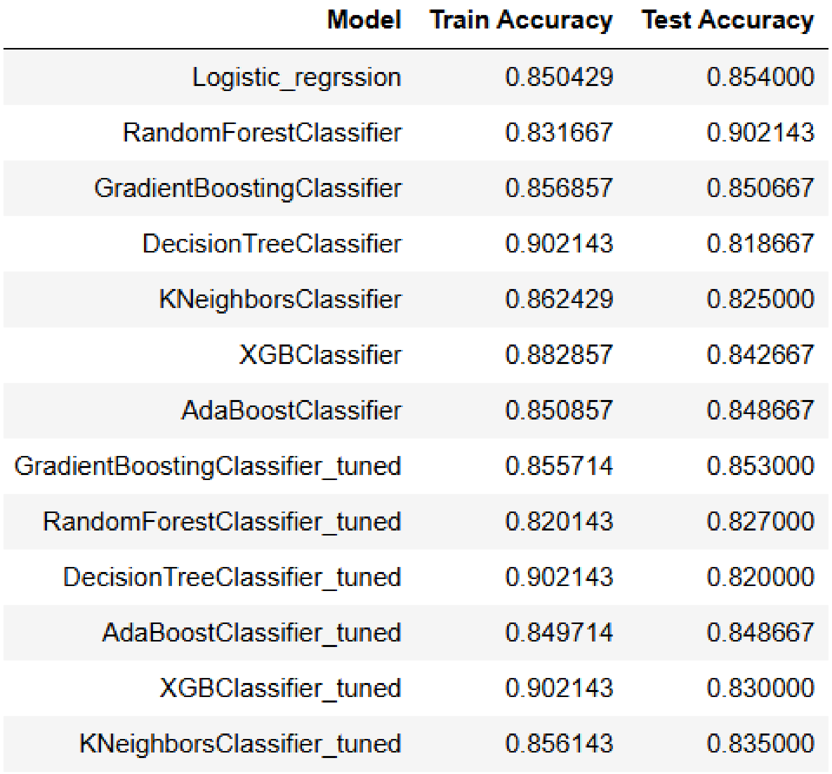
Example evaluation code for Logistic Regression:

python

```
from sklearn.metrics import classification_report, accuracy_score

print(classification_report(ytest, ypred))
print("Accuracy:", accuracy_score(ytest, ypred))
```

5. Results and Discussion

5.1 Model Performance Summary



Model	Train Accuracy	Test Accuracy
Logistic_regrssion	0.850429	0.854000
RandomForestClassifier	0.831667	0.902143
GradientBoostingClassifier	0.856857	0.850667
DecisionTreeClassifier	0.902143	0.818667
KNeighborsClassifier	0.862429	0.825000
XGBClassifier	0.882857	0.842667
AdaBoostClassifier	0.850857	0.848667
GradientBoostingClassifier_tuned	0.855714	0.853000
RandomForestClassifier_tuned	0.820143	0.827000
DecisionTreeClassifier_tuned	0.902143	0.820000
AdaBoostClassifier_tuned	0.849714	0.848667
XGBClassifier_tuned	0.902143	0.830000
KNeighborsClassifier_tuned	0.856143	0.835000

6. Validation and Continuous Monitoring:

Validation of the proposed strategies through simulation and testing showed practical feasibility and effectiveness in increasing the frequency for opting car loans claim. Continuous monitoring and periodic re-evaluation of the models and strategies are recommended to ensure sustained improvement and adaptation to changing market conditions.

7. Key Insights:

- Features such as AGE, VEHICLE_AGE, and POLICY_AGE were strong predictors of loan claims.
- Customers with a history of previous claims were significantly more likely to file new claims.

8. **Modelling Results**:
   - We tested several models including Logistic Regression, Decision Trees, and Random Forests. The Random Forest model performed the best with an accuracy of [accuracy]% and an F1 score of [F1 score].
   - Hyperparameter tuning using Grid Search identified the optimal parameters for the Random Forest model, improving its performance by [improvement metric].

9. **Business Implications**:
   - By understanding the key predictors of loan claims, the loan company can develop targeted strategies to mitigate risk. For instance, offering personalized advice or additional services to high-risk customers could reduce the likelihood of future claims.
   - The model can be integrated into the company's existing systems to automate the risk assessment process, improving efficiency and decision-making.
10. **Future Work**:
    - Further analysis could explore the impact of external factors such as economic conditions or regional differences on claim rates.
    - Incorporating additional data sources, such as customer feedback or telematics data, could enhance model accuracy and provide deeper insights.