Project 2 – Part 3 Presentation

Name : Natumanya Duncan

08/10/2023

Unsupervised Machine Learning

Stakeholder : Insurance Company

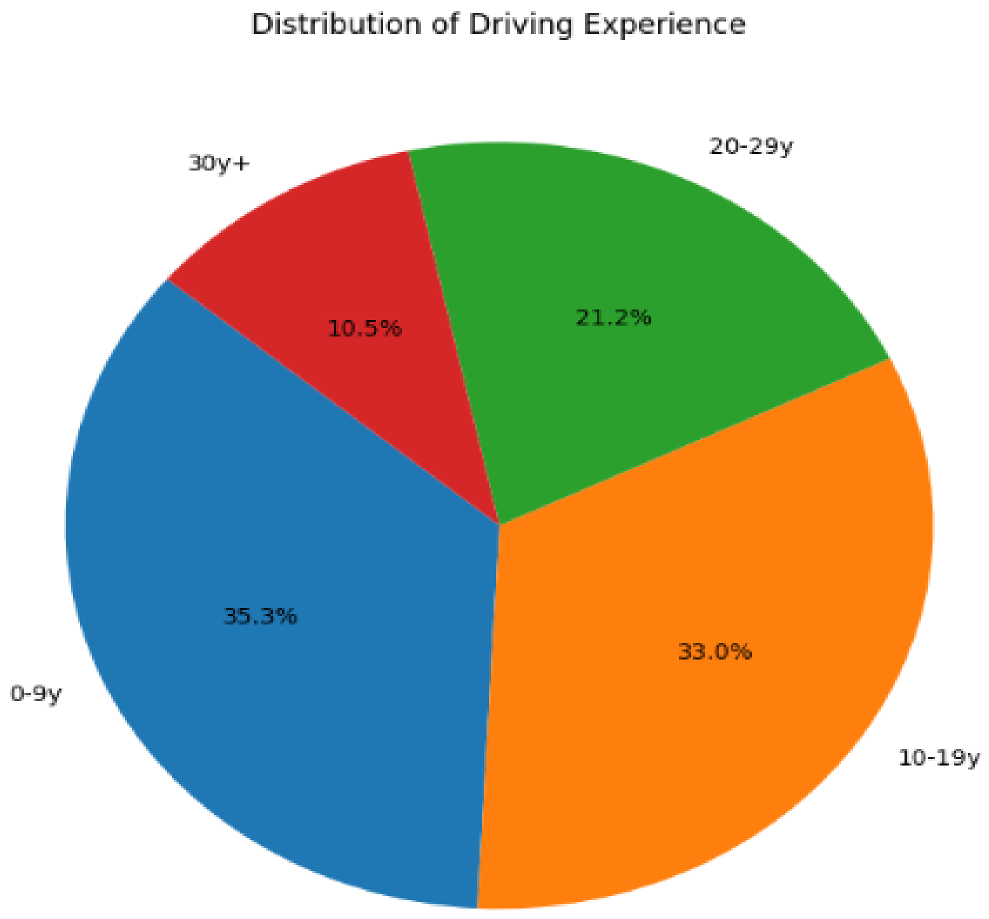
- Assessing risk associated with insuring individuals based on their characteristics and driving behavior.

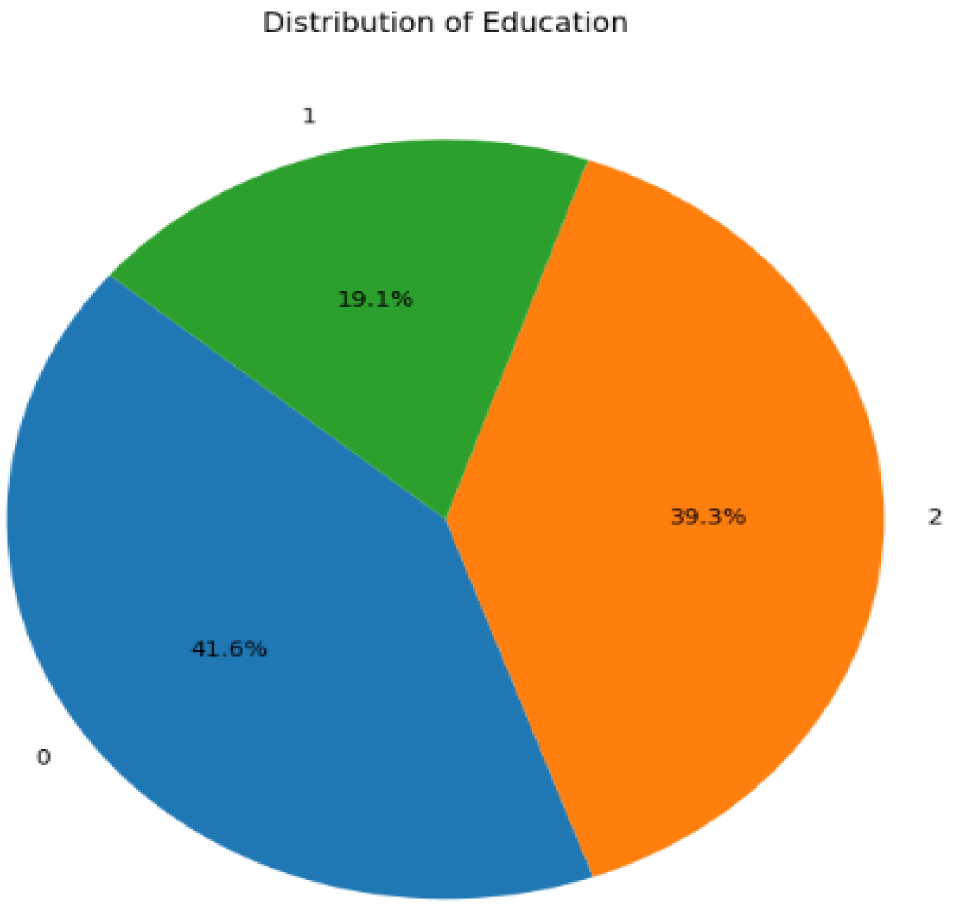
• Identifying factors that correlate with higher risk behavior (e.g., past accidents, DUIs, speeding violations) as well as factors that may influence risk (e.g., age, driving experience, credit score)

INTRODUCTION ABOUT THE DATA

- Demographic Information: Age, Gender, Race
- Driving Experience: Number of years of driving experience
- Education Level: Highest level of education attained
- Financial Details: Income, Credit Score
- Vehicle Details: Vehicle Ownership, Vehicle Year, Vehicle Type
- Marital and Family Status: Marital Status, Children
- Location Information: Postal Code
- Driving Behavior: Annual Mileage, Speeding Violations, DUIs, Past Accidents
- Outcome: This could potentially be an indicator or label used for assessing risk. Impact of False Positives and False Negative

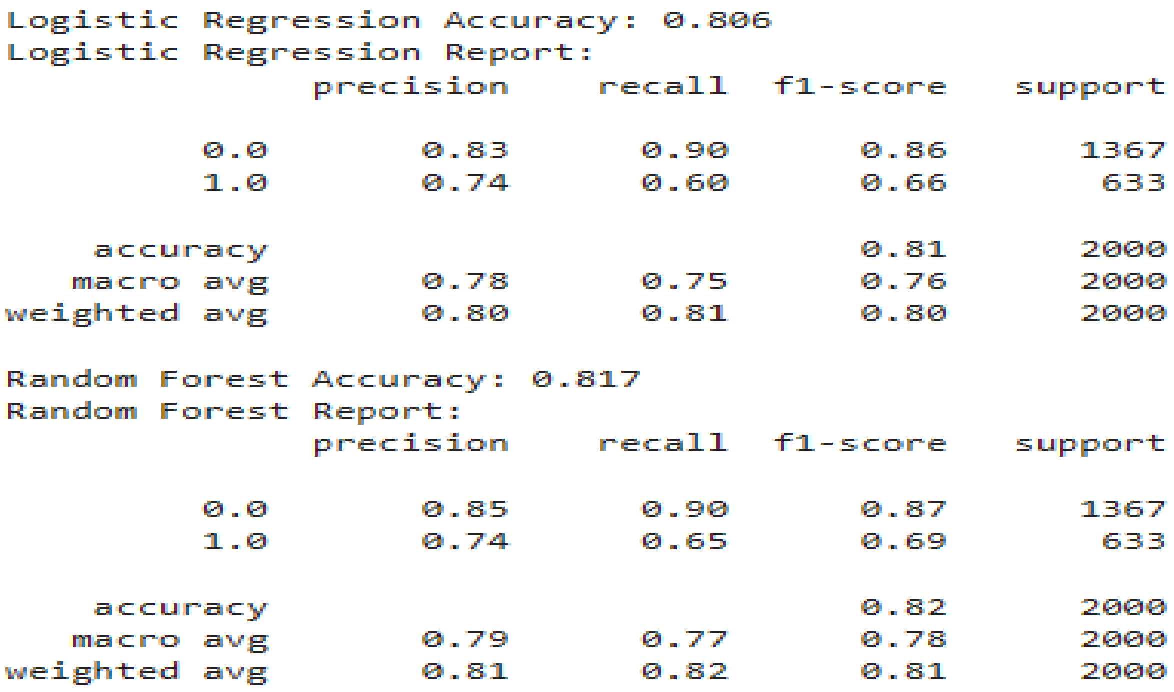
Some of the visuals about the data





Models Used

Summary of how the models performed;



```
Logistic Regression Accuracy: 0.806
Logistic Regression Report:
              precision    recall  f1-score   support

         0.0       0.83      0.90      0.86      1367
         1.0       0.74      0.60      0.66       633

    accuracy                           0.81      2000
   macro avg       0.78      0.75      0.76      2000
weighted avg       0.80      0.81      0.80      2000

Random Forest Accuracy: 0.817
Random Forest Report:
              precision    recall  f1-score   support

         0.0       0.85      0.90      0.87      1367
         1.0       0.74      0.65      0.69       633

    accuracy                           0.82      2000
   macro avg       0.79      0.77      0.78      2000
weighted avg       0.81      0.82      0.81      2000
```

Strengths and limitations of the model

- **Strengths**

• Rich Data Attributes: The dataset contains a diverse set of attributes including demographic details, driving behavior, education, income, and more. This richness in data allows for a comprehensive analysis of individuals and their potential risk profiles.

• Relevance to Insurance Risk Assessment: The dataset includes key factors that are highly relevant to insurance risk assessment, such as driving experience, past accidents, DUIs, and credit score. These features can significantly contribute to accurate risk evaluation.

• Large Dataset: With 10,000 rows, the dataset is of a substantial size, which can provide a robust foundation for training and validating a predictive model. A larger dataset generally leads to more reliable and stable models.

Cont’d

- **Limitations**
- Missing Values: There are missing values in the dataset, particularly in the 'CREDIT_SCORE' and 'ANNUAL_MILEAGE' columns. Depending on the approach taken to handle these missing values, it might introduce some level of uncertainty or bias into the model.

• Limited Historical Data: The dataset may not cover a long period of time. More historical data would provide a more comprehensive understanding of individual behavior and risk patterns.

• Lack of Contextual Information: The dataset might not include certain external factors that could impact risk assessment, such as weather conditions, traffic patterns, or specific driving routes.

RECOMMENDATIONS FOR STAKEHOLDERS

- **Prioritize False Positive Reduction**: Given the class imbalance and the potential cost associated with false positives, it is recommended to focus on reducing the occurrence of false positives. This can be achieved through techniques such as adjusting the model's decision threshold or employing additional post-modeling strategies to validate positive predictions
- **Regular Model Updates and Monitoring:** The model's performance should be periodically monitored and re-evaluated. As the business landscape evolves, it's crucial to ensure that the model remains accurate and reliable. Regular updates, possibly with new data or improved algorithms, will help maintain its effectiveness over time.