Exploring Car Insurance Trends

OUARDIGHI Omar

2023-12-27

Associated Insurance Problem:

The primary insurance problem this dataset can help address is understanding the factors that contribute to the likelihood of an insurance claim being filed. By analyzing these attributes, we can gain insights into risk factors that influence insurance claims. This is crucial for insurance companies for several reasons:

- Risk Assessment: Identifying high-risk drivers for more accurate policy pricing.
- Policy Personalization: Tailoring insurance policies based on individual risk profiles.
- Claim Prediction: Anticipating the likelihood of claims to better manage reserves and resources.
- Fraud Detection: Identifying patterns that may suggest fraudulent claims.

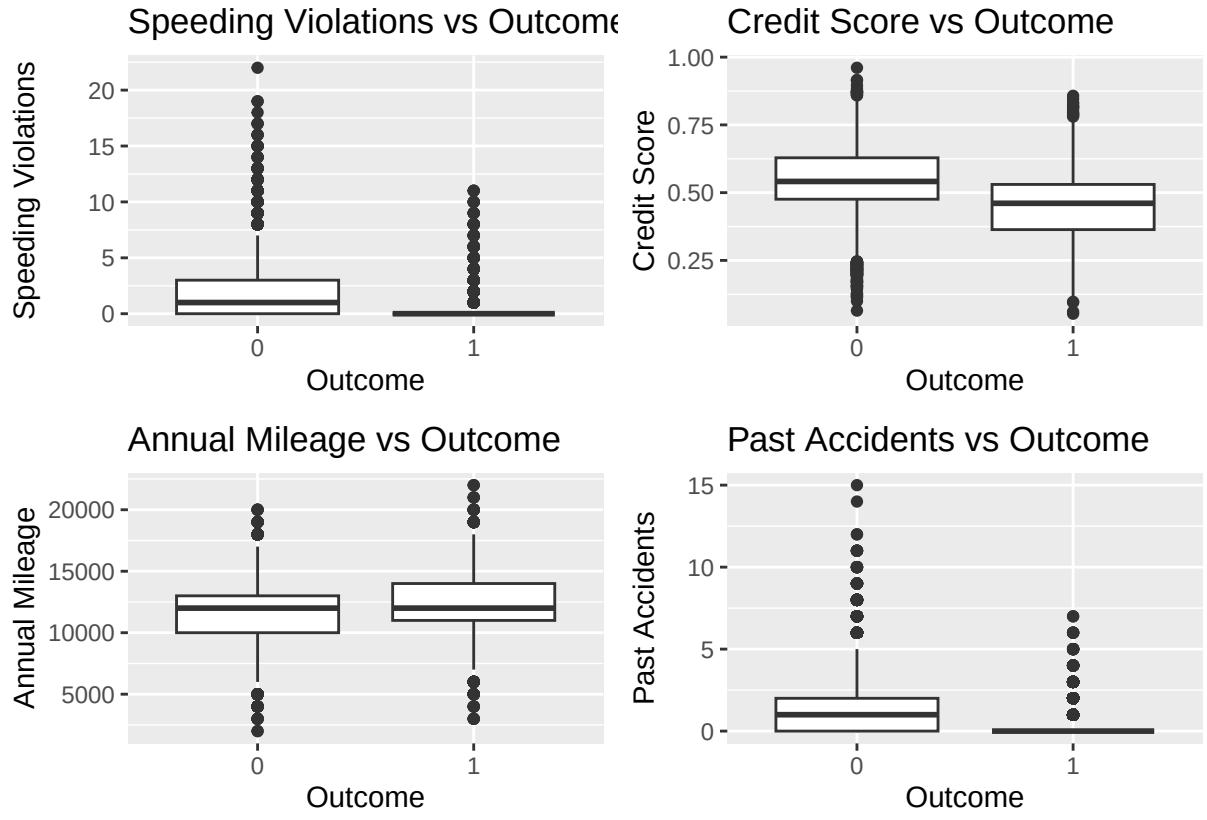
The outcome of this analysis could inform strategies for premium calculation, policy design, and overall risk management in the insurance sector.

Data Overview

- **Dataset Composition:** The dataset contains various attributes related to individual insurance policyholders and their driving history (10000 obs. and 19 variables)
- **Key Attributes:**
  - **ID:** A unique identifier for each policyholder.
  - **Demographics:** Includes age (AGE), gender (GENDER), and race (RACE).
  - **Driving Experience (DRIVING_EXPERIENCE):** Categorized by years of experience.
  - **Education Level (EDUCATION):** Indicates the highest level of education attained.
  - **Income Bracket (INCOME):** Classifies the policyholder's income.
  - **Credit Score (CREDIT_SCORE):** A numerical representation of the policyholder's creditworthiness.
  - **Vehicle Ownership (VEHICLE_OWNERSHIP)**: Indicates whether the policyholder owns a vehicle.
  - **Vehicle Year (VEHICLE_YEAR):** Categorizes the vehicle as either 'before 2015' or 'after 2015'.
  - **Marital Status (MARRIED) and Children (CHILDREN):** Provide family background.
  - **Postal Code (POSTAL_CODE):** Represents the geographical area of the policyholder.

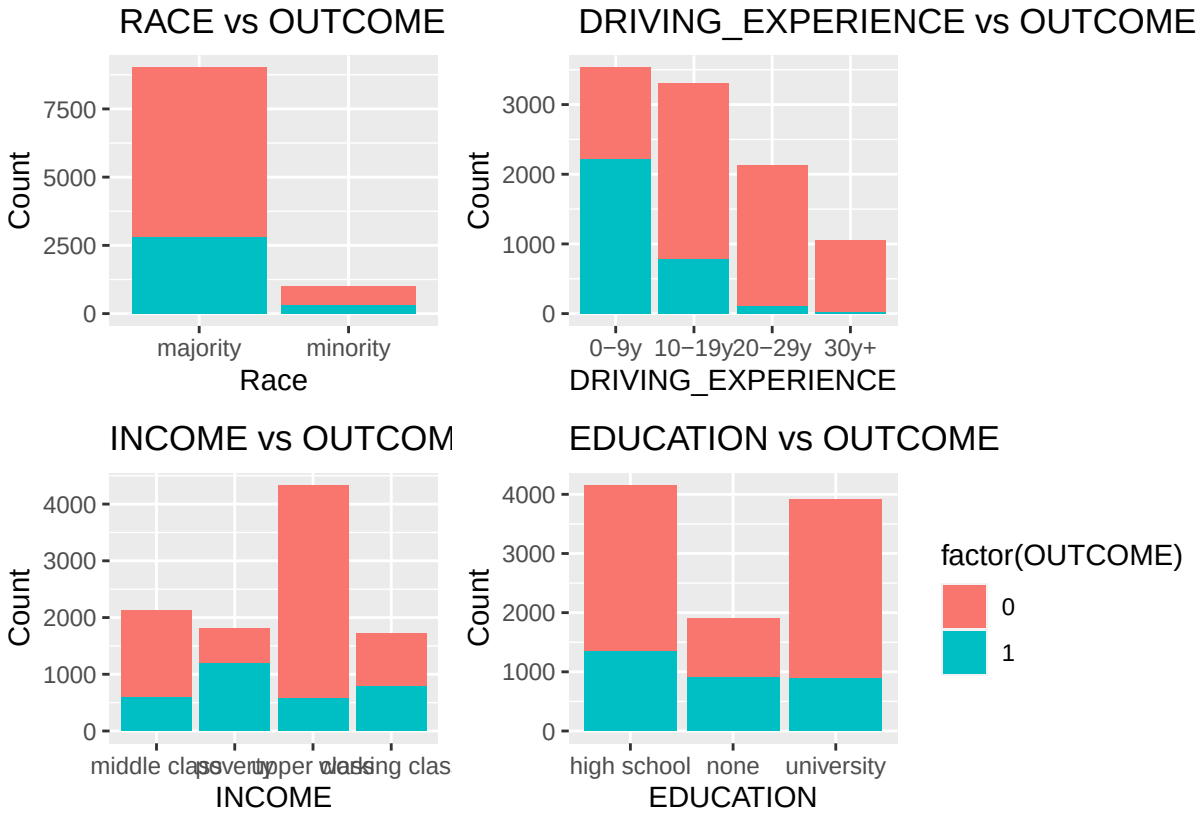
- **Annual Mileage (ANNUAL_MILEAGE):** The estimated number of miles driven per year.
- **Vehicle Type (VEHICLE_TYPE):** Type of vehicle insured.
- **Speeding Violations (SPEEDING_VIOLATIONS), DUIs (DUIS), and Past Accidents (PAST_ACCIDENTS):** Indicate driving history.
- **Outcome (OUTCOME):** Whether an insurance claim was filed (1) or not (0).

Exploratory Data Analysis



Boxplots reveal relationships between SPEEDING_VIOLATIONS, CREDIT_SCORE, ANNUAL_MILEAGE, and PAST_ACCIDENTS with the OUTCOME variable. Key findings:

- Speeding Violations: Differences in median and range suggest an impact on outcomes.
- Credit Score: Notable distributions indicate a potential influence on the outcome.
- Annual Mileage: Differences suggest a relationship between driving amount and outcomes.
- Past Accidents: Clear distinctions hint at the significance of accident history in outcomes.

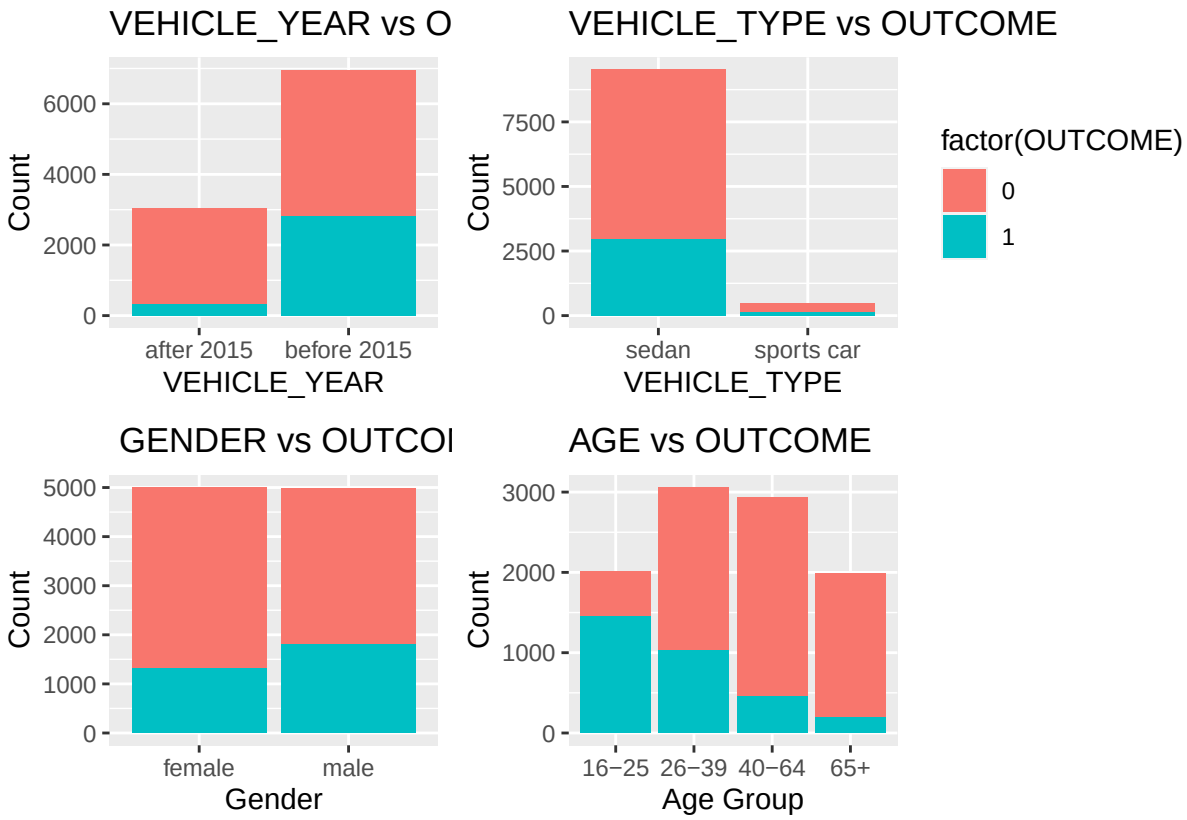


RACE: The difference in claim proportions between majority and minority races is small, with a marginally higher proportion of claims in the minority group.

DRIVING_EXPERIENCE: A clear trend is visible where less experienced drivers (0-9 years) have a higher proportion of claims, which decreases as driving experience increases.

EDUCATION: Individuals with no education have a higher proportion of claims compared to those with high school or university education.

INCOME: The 'poverty' income group has a noticeably higher proportion of claims, whereas the 'upper class' group has the lowest.



VEHICLE_YEAR: Owners of older vehicles (before 2015) have a higher proportion of claims compared to those with newer vehicles (after 2015).

VEHICLE_TYPE: Sports car owners have a higher proportion of claims than sedan owners.

AGE: The proportion of insurance claims is notably higher in the younger age group (16-25), and it decreases with age.

GENDER: The difference in the proportion of claims between males and females is subtle, but males show a slightly higher propensity for claims.

Modeling:

In order to predict the likelihood of insurance claims being filed, we used logistic regression and random forest models. The dataset undergoes one-hot encoding to handle categorical variables, ensuring compatibility with modeling techniques. The preprocessing steps, including the removal of an identifier variable and conversion of the outcome variable to a factor, contribute to data readiness.

Following a meticulous train-test split, logistic regression and random forest models are fitted to the training data. The logistic regression model utilizes the generalized linear model framework, while the random forest model leverages an ensemble of decision trees. Both models are well-suited for binary classification tasks and are implemented using R packages, namely **`glm`** and **`randomForest`**.

Logistic Regression

The logistic regression model exhibits robust predictive capabilities, with high accuracy, sensitivity, specificity, and precision. The AUC value further confirms its effectiveness in distinguishing between positive and

negative outcomes.

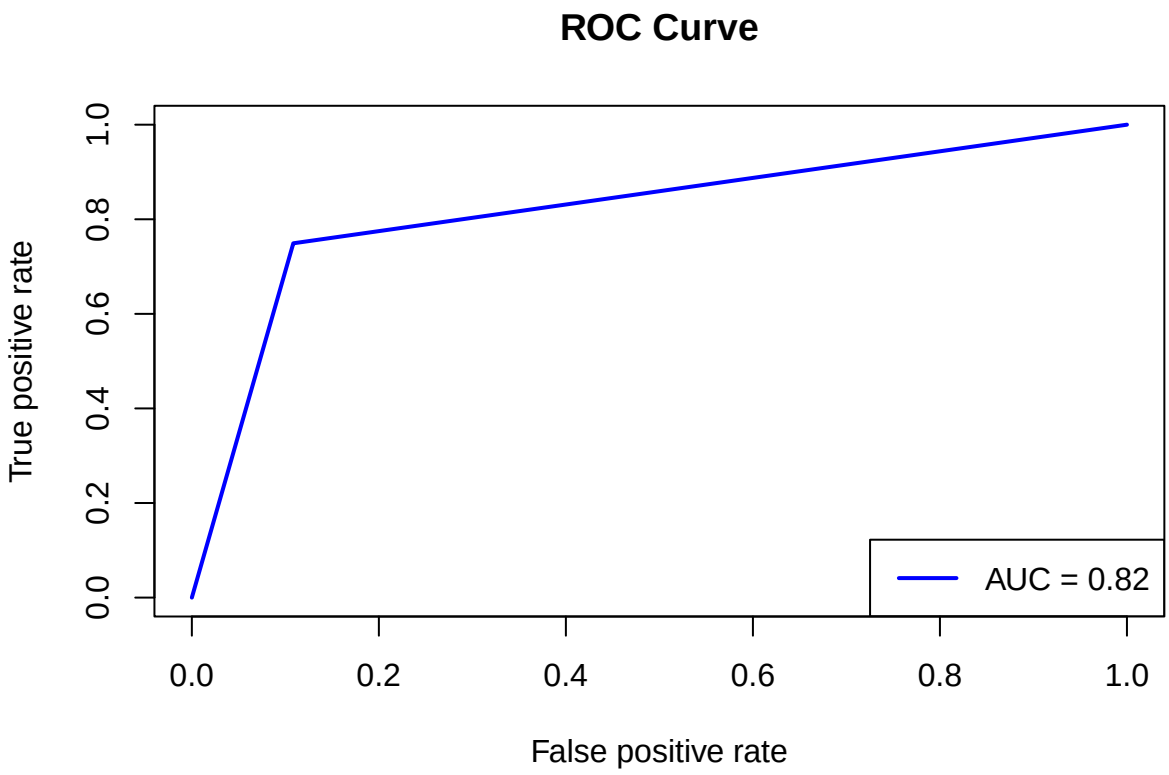
```
## Accuracy: 0.8469235
```

```
## Sensitivity: 0.8914785
```

```
## Specificity: 0.7492013
```

```
## Precision: 0.8863143
```

```
##            Reference
## Prediction    0    1
##          0 1224  157
##          1  149  469
```



Random Forest Classifier:

The random forest classification model, with 500 trees and 5 variables considered at each split, exhibits robust performance:

```
## Accuracy: 0.8374187
```

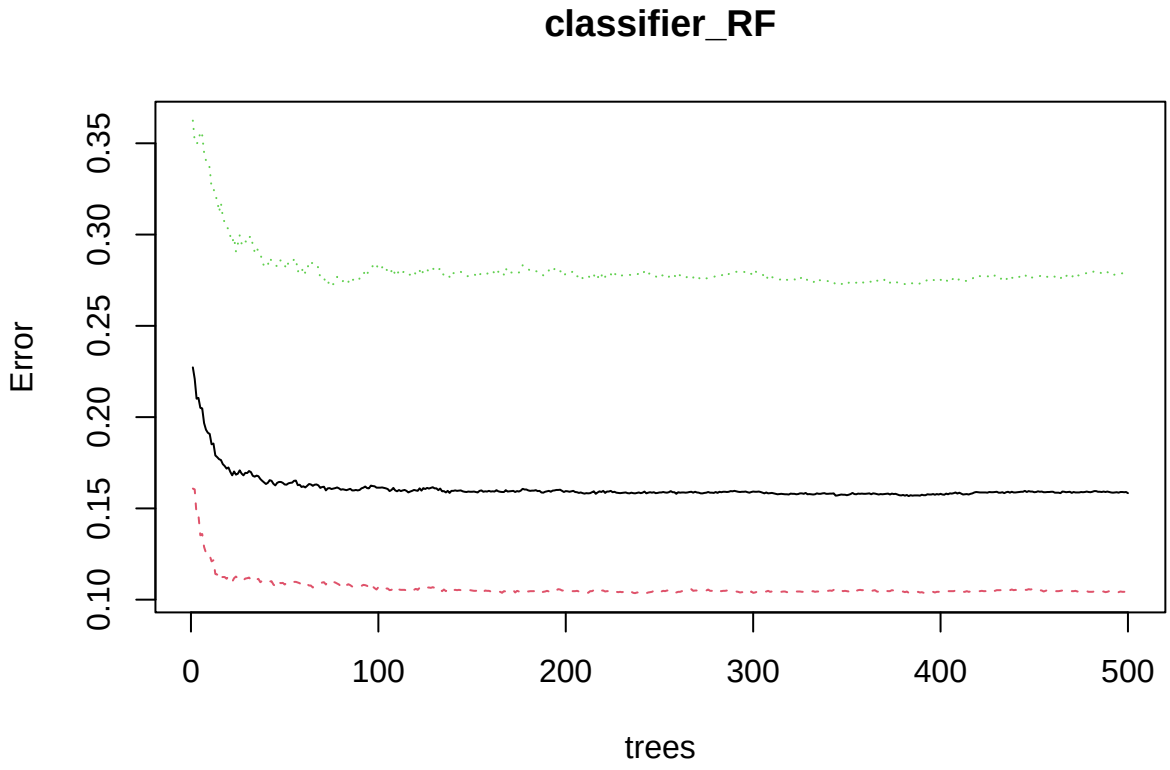
```
## Sensitivity: 0.8790969
```

```
## Specificity: 0.7460064

## Precision: 0.8836018

##            Reference
## Prediction    0    1
##          0 1207  159
##          1  166  467
```

The Plot shows that the Error rate is stabilized with an increase in the number of trees.



Analysis and Conclusion

- **Performance**: Both models show good performance, with logistic regression slightly outperforming random forest in terms of accuracy and sensetivity for claims.
- **Model Choice**: Logistic regression, despite its simplicity, performs competitively with the more complex random forest model in this case. This might be due to the nature of the data or the way features interact in this particular dataset.
- **Potential Improvements**: Further tuning of hyperparameters, feature engineering, or trying other models like gradient boosting might improve performance. Additionally, investigating and understanding feature importance could provide insights for better model training and interpretation.

In summary, for this dataset, logistic regression is a strong candidate given its performance and simplicity, but there's room for exploration with more complex models or further data analysis.