Carleton University

Age vs. Experience: Which Factor Has a Greater Influence on Accidents?

Exploring the Impact of Age and Experience on Accident Prediction

Sicheng (Kevin) Lu

Personal Projects # 1

Bachelor of Economics Honours

Minor in Statistics

Introduction

- For the road safety, it is crucial to understand the factors that contribute to accident as effective prevention strategies.
- Driving Age and Experience shows debated as key influences on driver’s likelihood of being involved in an accident.
- Question:
  - Is it wisdom and caution that produce age or confidence and skills gained through experience that play a larger role in accident prevention?

Introduction (Continued)

- Will find out by exploring to see if age and experience prove to be significant on accident prevention.
- Examine data trends and statistical relationships.
- Provides valuable insights for policymakers, insurers, and divers alike in shaping more effective safety protocols and policies.

Where did I get a Dataset??

- I received a dataset from Kaggle's website.
  - Kaggle is a data science and artificial intelligence platform.
  - Contains with many kind of datasets
  - Published by large companies and organizations

  Datasets: dataset_traffic_accident_prediction1

  Data link: https://www.kaggle.com/datasets/saurabhshahane/road-traffic-accidents/data

Data Cleaning with MySQL

- **Data Cleaning can be done by MySQL Workbench**
- **Found some of the Data are missing**
  - **Some of them have blank data**
  - **Adjust the blank data to let say assume as 0 or "unknown"**

Data Cleaning with MySQL (Continued)

- Found Six Columns have left some blank data
- Here the missing Data and how it can fix:
- Accident_Severity data, presume it is “Unspecified”
- Road_Condition and Road_Type, let say “Not Determined”
- Weather and Time_of_Day, would say replaced from missing data to “Unknown”
- Traffic_Density would be “0”

Data Cleaning with Excel

- Once the data is all cleaned with MySQL Workbench
- Saved the cleaned data into CSV file
- Opened the cleaned data CSV file via Excel
- But...
- What happened next?

Data Cleaning with Excel

- All the data were in only One Cell.
- How to fix it???
- By splitting the text into columns
- Fixed all the data by Column Wizard
- When it's fixed, all the data is now organized by splitting text into columns

Data Analysis with Python

- Independent Variables: ‘Driver_Age’ , ‘Driver_Experience’
- Dependent Variable: ‘Accident’

Figure 1: Scatter Plot: Driving Age vs Driver Experience

Figure 1: Scatter Plot: Driving Age vs Driver Experience (Continued)

- Clear positive correlation between Driver Age and Driver Experience
- Age increases, more driver experience
- But.....
- Data spread out due to many variability in driver experience at any given age
- While accident in occurrence, not strongly associated with either driving age or driver experiences. It can happen vary.
- Possibilities Alcohol or speeding as other factors might likely play an important role determine accident risk.

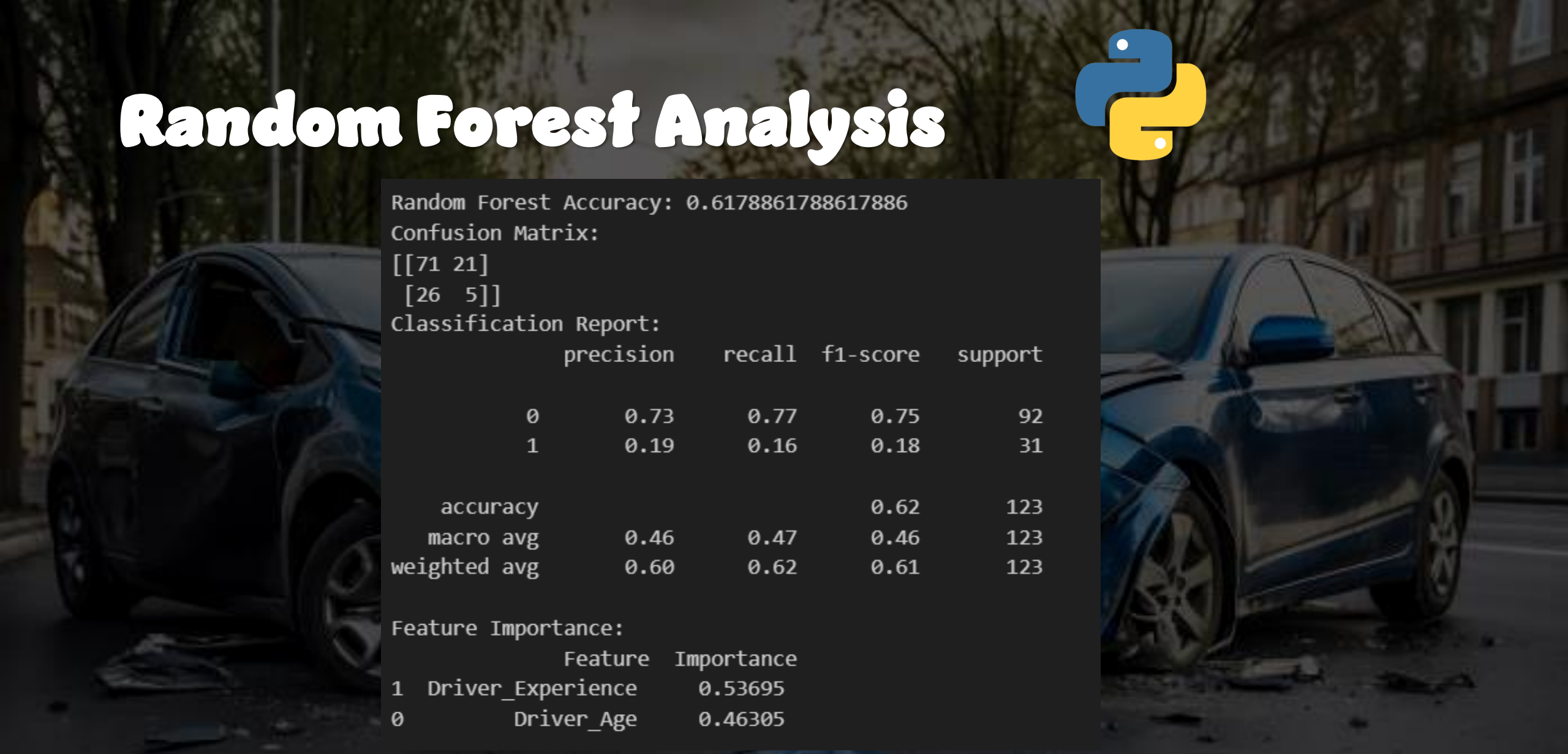
Random Forest

Why did I choose Random Forest???

- Dependent Variable “Accident” considered as binary classification (0 and 1)
- Non-linear relationships
- Reduces overfitting through:
  - Bagging
  - Robot to noise
  - Missing data
  - Multicollinearity

- Works well with imbalanced data
- Generates probability estimates for predictions

Random Forest Analysis



```
Random Forest Accuracy: 0.6178861788617886
Confusion Matrix:
[[71 21]
 [26  5]]
Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.77      0.75        92
           1       0.19      0.16      0.18        31

    accuracy                           0.62       123
   macro avg       0.46      0.47      0.46       123
weighted avg       0.60      0.62      0.61       123

Feature Importance:
              Feature  Importance
1  Driver_Experience     0.53695
0         Driver_Age     0.46305
```

Figure 2: Random Forest Accuracy Output

Random Forest Analysis (Continued)

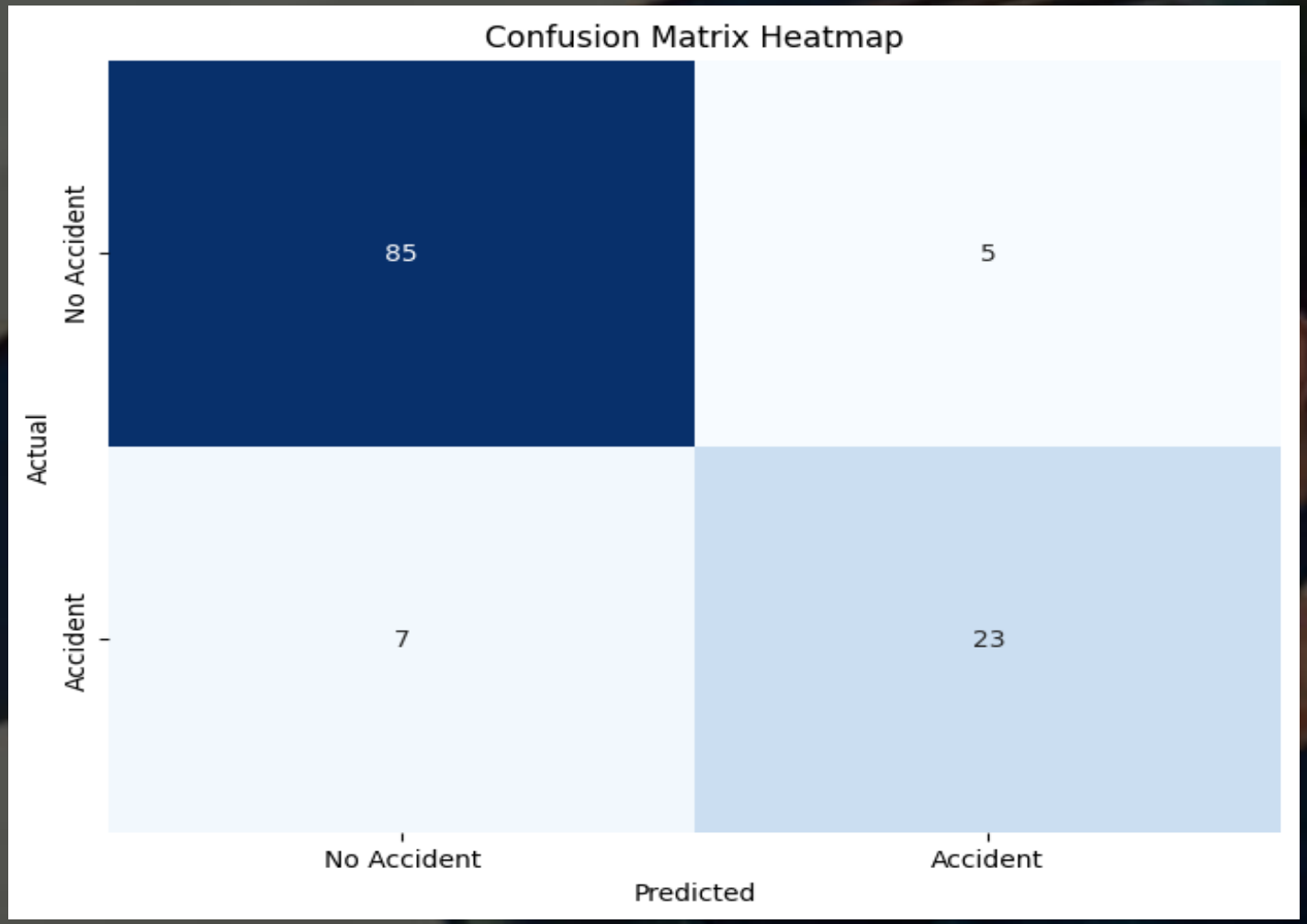
- Accuracy = 62% (0.62)
  - Moderate performance, suggest for room for improvement
- Confusing Matrix:
  - True Negatives = 71 (Correctly reported as “No Accident”)
  - False Positives = 21 (Incorrectly reported as “Accident” but with “No Accident”)
  - False Negatives = 26 (Missed “Accidents”)
  - True Positives = 5 (Correctly predicted “Accidents”)

Random Forest Analysis (Continued)

- Classification Report:
  - Precision: Class 0 (No Accident) = 73% (good), Class 1 (Accident) = 19% (poor)
  - Recall: Class 0 = 77% (high), Class 1 = 16% (low)
  - F1 Score: Class 0 = 0.75 (solid), Class 1 = 0.18 (weak)
  - Macro Avg: Both classes = 46% (weak)
  - Weighted Avg: 61% better for “No Accident”
- Feature Importance:
  - Driver Experience: 53.7%
  - Driver Age: 46.3%

  Driver Experience are more influential.

Confusion Matrix Heatmap



- Predictions whether an accident will occur or not
- 85 Cases truly reported as “No Accident”
- 23 Cases truly reported as “Accident”
- 7 Cases mistakenly reported as “No Accident” even though supposed to report as “Accident”
- 5 Cases mistakenly reported as “Accident” even though there was “No Accident”

Figure 3: Confusion Matrix Heatmap for Accident

Confusion Matrix Bar Plot

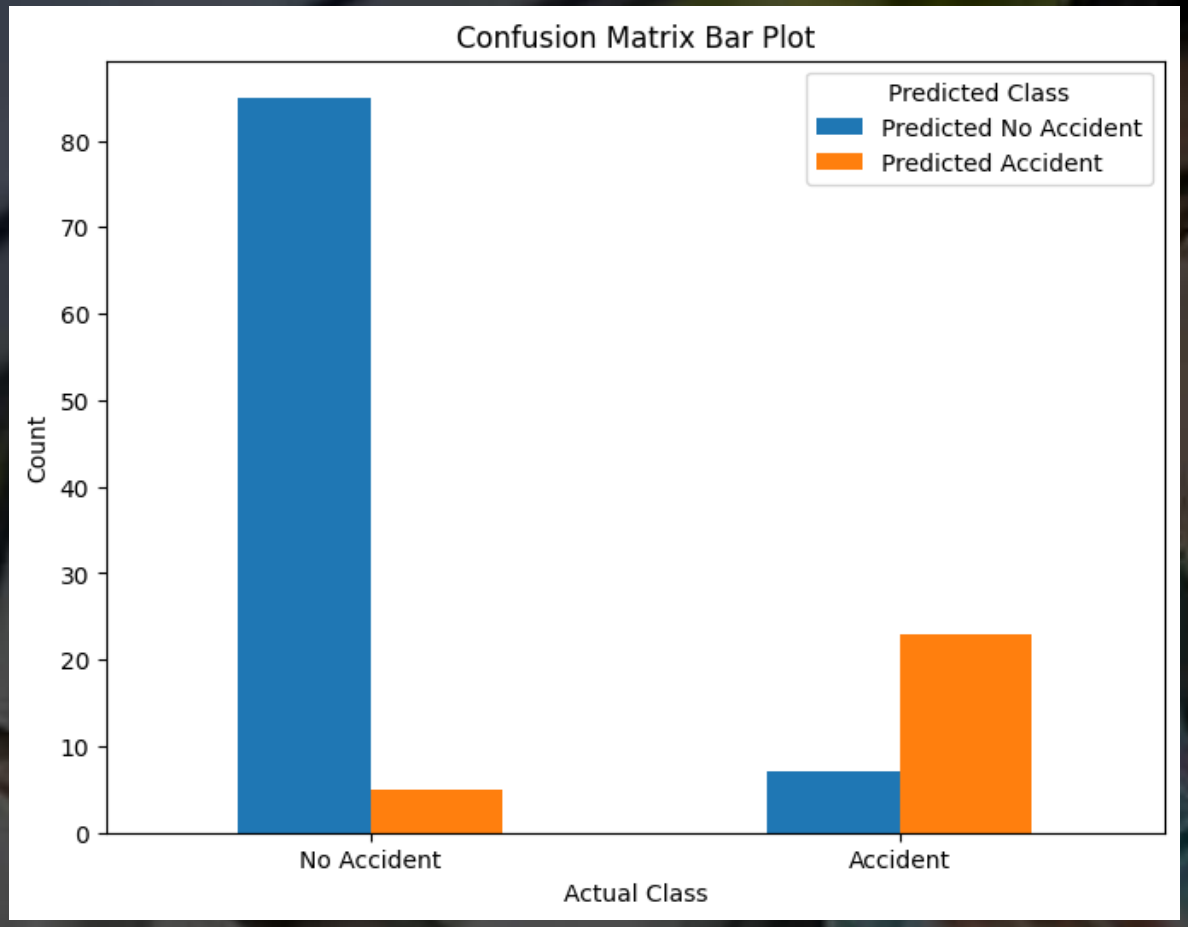


Figure 4: Confusion Matrix Bar Plot for Accident

- Horizontal Axis = True Classes "No Accident" and "Accident"
- Vertical Axis = # of instances (Count)
  - Blue Bar = "No Accident"
  - Orange Bar = "Accident"
  - Correctly placed as "No Accident" in the taller blue bar
- Correctly placed as "Accident" in the taller orange bar
- Mistakenly placed "Accident" in shorter orange bar
- Mistakenly placed "No Accident" in shorter blue bar

Driver Age vs Accident & Driver Experience vs Accident

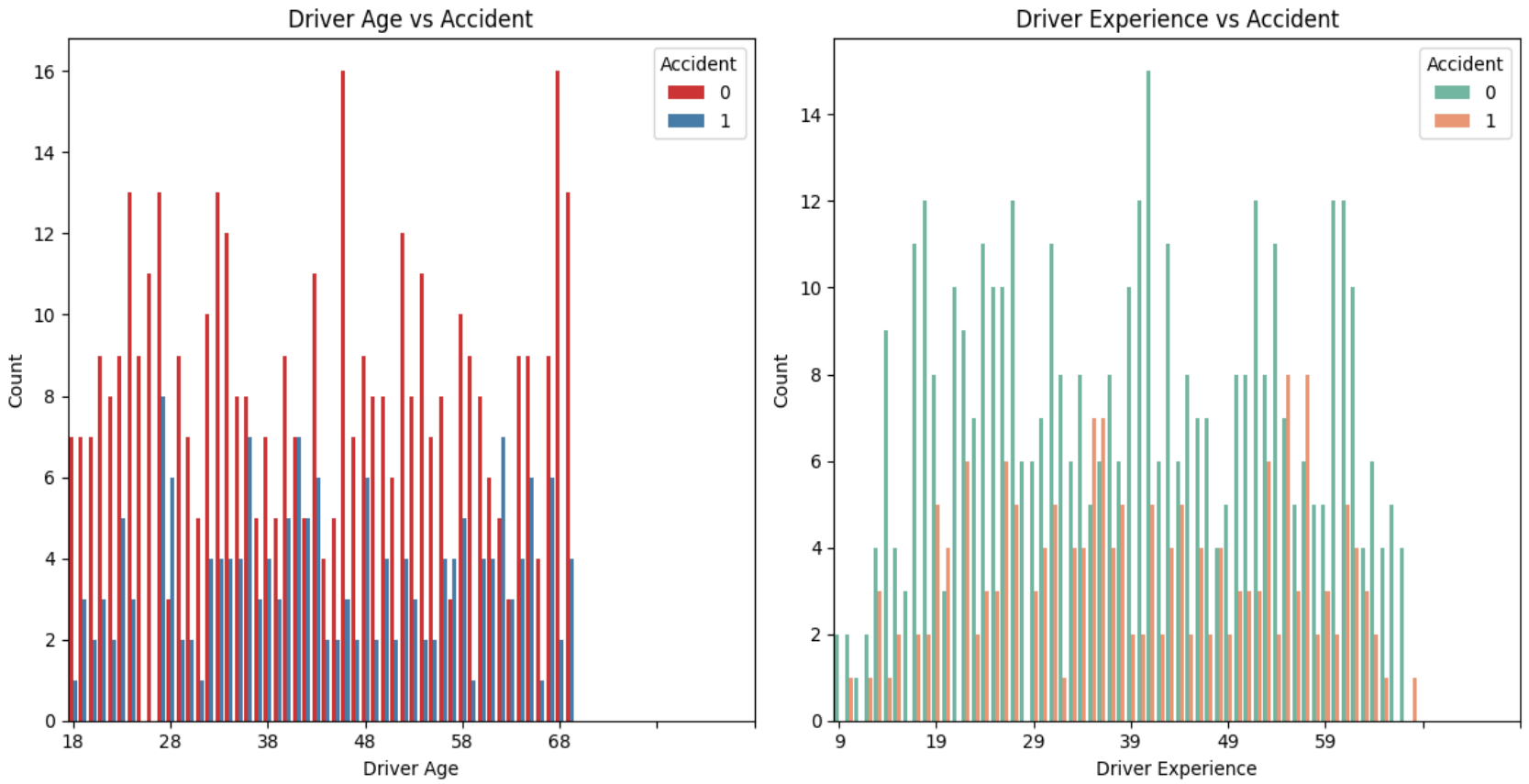


Figure 5: Driver Age vs Accident and Driver Experience vs Accident

Driver Age vs Accident

- Driver Age vs Accident (Graph in the left side)
  - Red Bars = No Accident (Accident = 0)
  - Blue Bars = Accident (Accident = 1)
  - X – Axis = Driver Age (Age range = approx. 18 – 68)
  - Y – Axis = Count (# of drivers in each age group)
  - Most Drivers did not experience accidents across in all age ranges (Red Bars)
  - Drivers with accidents have fewer in number but distributed across all ages (Blue Bars)
  - No clear age group with significantly higher accident rates

Driver Experience vs Accident

- Driver Experience vs Accident (Graph on the ride side)
- Green Bars = No Accident (Accident = 0)
- Orange Bars = Accident (Accident = 1)
- Drivers with no accidents dominate across all experience levels (Green Bar)
- Accident shows relatively infrequent and scattered across experience groups
- No strong trends suggest that experience consistently affects accident likelihood.

Results from Driving Age and Driving Experience vs Accident Graph

Non of driver age and experience shows strong correlation with accidents because most drivers avoid accident in regardless of these factors.

Correlation Heatmap: Driver Age, Driver Experience, and Accident

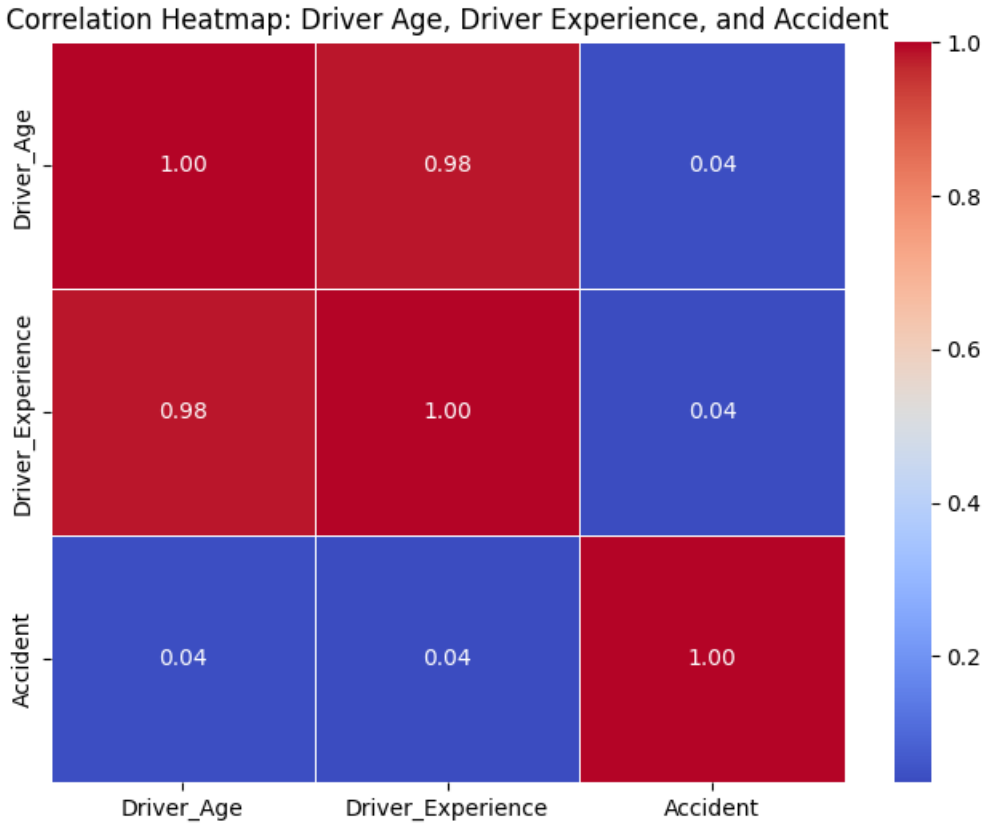


Figure 6: Correlation Heatmap: Driver Age, Driver Experience, and Accident

- Driving Age vs Experience:
  - Correlation = 0.98 (Very High)
  - Older drivers may have more experienced
- Driving Age vs Accident:
  - Correlation = 0.04 (Very Low)
  - Driver Age has almost no relationship with accident occurrence
- Driver Experience vs Accident:
  - Correction = 0.04 (Very Low)
  - Experience shows negligible influence on accidents

Both Driver Age and Driver Experience are not significant impact to show an evidence of an accidents, suggest other factors may play a larger role in accidents. However, strong correlation between age and experience is intuitive and expected.

Conclusion

- Primary Insights: Driver Age and experience shows negligible direct impact on accident likelihood as according to the low correlation values (0.04)
- Age vs Experience have stronger correlation (0.98) where when driver gets older, more experiences
- Findings:
  - Accident happened randomly across all age and experience without clear patterns
  - Other factors like alcohol consumption or speeding as these are more likely to have significant roles in accidents

Conclusion (Continued)

- Model Performance:
  - Random Forest Model = 62%
    - Struggled predicting accidents accurately due to low precision and recall for accidents
  - Driver Experience (53.7%) are more influential than driver age (46.3%) in predicting accidents.

Age and experience along are insufficient predictors of accidents, underscoring the need to investigate additional factors, maybe alcohol consumption or speeding which might helps to improve road safety measure.