CS777 – Term Project Proposal - Airlines

**Student: Yuhan Xu**

1. Data set description: Provide a detailed description of the public data set you have selected, including its source, format, and any relevant details about the data.

|  |
| --- |
| *Overview:*  *The "Airline Safety" dataset provides a comprehensive overview of various safety metrics related to different airlines, with a particular focus on incidents, fatal accidents, and fatalities. The data spans over three decades, from 1985 to 2014, and is intended to facilitate an exploration into the historical safety records of airlines. This dataset was utilized in the analysis presented in the story "Should Travelers Avoid Flying Airlines That Have Had Crashes in the Past?".*  *Source:*  *Aviation Safety Network*  *Key Features:*  *airline: The name of the airline. An asterisk (\*) indicates that data includes regional subsidiaries.*  *avail\_seat\_km\_per\_week: Represents the available seat kilometers flown every week, providing a measure of the airline's size and the extent of its operations.*  *incidents\_85\_99: The total number of incidents recorded by the airline from the year 1985 to 1999.*  *fatal\_accidents\_85\_99: The total number of fatal accidents that occurred with the airline from the year 1985 to 1999.*  *fatalities\_85\_99: The total number of fatalities that were recorded by the airline from the year 1985 to 1999.*  *incidents\_00\_14: The total number of incidents recorded by the airline from the year 2000 to 2014.*  *fatal\_accidents\_00\_14: The total number of fatal accidents that occurred with the airline from the year 2000 to 2014.*  *fatalities\_00\_14: The total number of fatalities that were recorded by the airline from the year 2000 to 2014.*  *Potential Use-Cases:*  *Safety Analysis: Analyzing the safety performance of different airlines over the specified periods.*  *Risk Assessment: Evaluating the risk factors associated with various airlines based on historical incidents and fatalities.*  *Trend Analysis: Identifying trends in airline safety and incidents over time.*  *Comparative Analysis: Comparing the safety metrics of different airlines to determine relative safety standings.* |

1. Research question: Clearly define your research question and explain why studying is important. What do you want to learn from the data?

|  |
| --- |
| *Main Question:*  *How do historical incidents, fatal accidents, and fatalities correlate with the size and extent of airline operations, and can we predict future safety metrics based on this historical data?*  *Importance:*  *Understanding the correlation between the historical safety metrics and the operational size of airlines can provide insights into the effectiveness of safety protocols and regulations over time. This study aims to discern patterns or trends in airline safety and potentially predict future safety metrics, aiding regulatory bodies, airline operators, and travelers in making informed decisions.* |

1. Machine Learning model: Specify the type of machine learning model you plan to use, such as classification or clustering, and explain why you have chosen this model.

|  |
| --- |
| *Model Type:*  *Regression Analysis*  *Rationale:*  *Given that the dataset contains numerical and time-series data, a regression model will be apt to analyze the relationships between different safety metrics and the operational size of airlines. Additionally, regression analysis can help predict future safety metrics based on historical data, providing valuable insights into potential future trends or risks in airline safety.* |

1. Expected outcomes: What do you expect to achieve after implementing your learning model? What do you hope to learn or discover from your data analysis?

|  |
| --- |
| *Upon implementing the regression model, the expectation is to:*  *Identify significant correlations between airline operational size and various safety metrics.*  *Determine if the historical safety metrics can reliably predict future metrics.*  *Uncover any notable trends or patterns in airline safety over the observed periods.*  *Provide a predictive model that can estimate future safety metrics based on given variables.* |

1. Evaluation plan: Explain how you plan to evaluate your project and assess the correctness of your model. What metrics or methods will you use to evaluate the effectiveness of your learning model? How well do you expect the model to work, and how will you measure its performance?

|  |
| --- |
| **Evaluation Metrics:**  R-Squared Value: To evaluate the goodness of fit of the regression model.  Root Mean Squared Error (RMSE): To understand the model’s prediction error.  Residual Analysis: To ensure the residuals are normally distributed and homoscedastic.  F1 Score: To balance the precision and recall, providing a metric that considers both false positives and false negatives.  Precision: To evaluate the accuracy of the positive predictions.  Recall: To determine the model's ability to capture and predict all relevant cases.  Confusion Matrix: To visualize the model’s predictions in terms of true positive, true negative, false positive, and false negative.  **Assessment Plan:**  Data Split: The dataset will be split into training and testing sets to validate the model's predictive power.  Cross-Validation: Employ cross-validation to assess the model's robustness and to avoid overfitting.  Comparison with Baseline Model: Compare the regression model with a baseline model to ensure it provides additional predictive power.  Residual Plots: Analyze residual plots to check for any patterns, ensuring that the model is not missing any information.  Precision-Recall Curve: Analyze the trade-off between precision and recall for different probability thresholds.  ROC Curve: Evaluate the trade-offs between true positive rate and false positive rate.  **Performance Expectation**:  The model is expected to provide a reasonable prediction of future safety metrics and uncover significant correlations between the variables. The inclusion of F1, precision, and recall, along with the confusion matrix, will provide a more nuanced understanding of the model's performance, especially in terms of its predictive capabilities and reliability in various scenarios. The accuracy and reliability of predictions will be critically assessed and transparently communicated, considering the potential implications on public perception and airline operations. |