Keep to one page

| Name | Joel Quek |
| --- | --- |
| Cleaning Approach | Removed rows with missing values in key delay columns and dropped features not available before the flight (e.g., actual delay values).  Handled infinite values by replacing them with NaN and subsequently dropping rows with NaN values in the 'ARR\_DELAY' column.  Standardized and encoded categorical features for model input. |
| Feature Engineering Done | Extracted time-based features like day\_of\_week, month, hour\_of\_day from the FL\_DATE column.  Created interaction terms such as distance\_airline\_interaction to capture airline-specific trends related to delay.  One-hot encoding was applied for categorical variables (AIRLINE, ORIGIN, DEST) for logistic regression. |
| Predictor Variables | CRS\_DEP\_TIME, DISTANCE, AIRLINE, ORIGIN, DEST, hour\_of\_day, day\_of\_week, and distance\_airline\_interaction. |
| Response Variable | Created a binary target variable, Delayed, where delays were defined as flights departing after the scheduled departure time (DEP\_DELAY > 0). |
| Model Pipeline(s) | **Random Forest**: Achieved 76.5% accuracy. The model performed well, but improvements are planned through hyperparameter tuning.  **SVM**: Reached 74.8% accuracy. Evaluated using AUC-ROC curves and classification metrics, showing room for further optimization.  Both models were evaluated using cross-validation to avoid overfitting and included metrics like precision, recall, and F1-score. |
| Improvements to be made | Further hyperparameter tuning for both models to optimize performance.  Experiment with additional models, such as XGBoost, to improve class imbalance handling.  Explore additional features (e.g., weather data, airport congestion) to enhance model predictions. |
| Anything Else? | Future work will involve further balancing the dataset and exploring different model architectures, including ensemble methods like Gradient Boosting. |