FLIGHT DELAY PREDICTION MODEL USING MACHINE LEARNING

A Report submitted for fulfillment of two week internship program
By

VADLAMUDI NIKHITHA CHOWDARY
21955A0413

Chief mentor: **Dr.Ch.V. Rama Padmaja**
Associate professor
Co-ordinator: **B.Indu**
Asst professor

CAREER DEVELOPMENT CENTRE
INSTITUTE OF AERONAUTICAL ENGINEERING
Dundigal, Hyderabad, Telangana 500043

MAY 2023

CERTIFICATE

This is to certify that the project report entitled **FLIGHT DELAY PREDICTION MODEL** submitted by **VADLAMUDI NIKHITHA CHOWDARY** to the Institute of Aeronautical Engineering, Telangana, in partial fulfillment for the two week internship program is a bonafide record of project work carried out by her under our supervision from May 22, 2023 to June 4,2023

DECLARATION

I, VADLAMUDI NIKHITHA CHOWDARY (21955A0413), hereby declare that the Project Report entitled "FLIGHT DELAY PREDICTION USING MACHINE LEARNING done by me under the guidance of Dr Ch. V. RAMA PADMAJA, is submitted in fulfillment of Two Weeks Internship Program

DATE: 4/06/2023 V.NIKHITHA

SIGNATURE OF THE CANDIDATE
PLACE: IARE,DUNDIGAL

ACKNOWLEDGMENT

I am pleased to acknowledge our sincere thanks to Board of management of IARE for their kind encouragement in doing this 2week internship and for completing it successfully. I am grateful to them. We convey our thanks to Dr. G RAMU and Dr.CH V RAMA PADMAJA for providing us the necessary support and details at the right time during the progressive reviews. We would like to express our sincere and deep sense of gratitude to our Project Guide Dr.CH V RAMA PADMAJA for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my internship

ABSTRACT

Flight delays can be a major inconvenience for both passengers and airlines, leading to financial losses and disruptions in travel plans. Accurate prediction of flight delays can help airlines and passengers make informed decisions, mitigate potential issues, and improve overall operational efficiency. In this study, we propose a data-driven approach for predicting flight delays using machine learning techniques. Our model leverages historical flight data, including factors such as departure time, weather conditions, airport congestion, and previous flight delays, to train and validate predictive algorithms. We extract relevant features from the data and employ advanced machine learning algorithms, such as random forest or gradient boosting, to develop a robust prediction model

Contents

1 Introduction (1)

- 1.1Motivation
- 1.2 Problem statement
- 1.3 Report Organization

2 Related Work (2-4)

- 2.1 Existing Methods
- 2.2 Proposed Solution
- 2.3 Software

3 Methodology (5-6)

4 Results (7-8)

- 4.1 result
- 4.2 graph

5 Conclusions (9-10)

TABLE OF FIGURES/GRAPHS

- result graph (8)

TABLE OF TABLES

- Dataset(5)

Acronyms

RNN- Recurrent Neural Networks
RFC- Random Forest Classifier
ROC AUC-Area under the ROC Curve

INTRODUCTION

1.1 Motivation

Flight delays pose significant challenges for the aviation industry, affecting both airlines and passengers. Delays result in increased costs, operational inefficiencies, and passenger dissatisfaction. Therefore, there is a pressing need to develop accurate and reliable flight delay prediction models. By leveraging advanced machine learning techniques and historical data, these models can provide valuable insights into the factors contributing to delays. The motivation behind this research is to improve operational efficiency, optimize resource allocation, and enhance the overall travel experience by developing a robust flight delay prediction model.

1.2 Problem statement

The unpredictability of flight delays presents a complex problem for the aviation industry. Traditional methods of estimating delays based solely on historical averages or expert judgment are often insufficient. The challenge lies in identifying the key factors that influence delays and developing a predictive model that can accurately forecast delays in real-time. This research aims to address this problem by leveraging machine learning algorithms to analyze a comprehensive set of features, including weather conditions, airport congestion, and historical delay data. By doing so, the aim is to develop a reliable flight delay prediction model that outperforms existing methods and provides actionable insights for airlines and passengers.

1.3Report Organization

This report is structured as follows: In the next section, The existing literature on flight delay prediction model is reviewed and highlights the gaps that this research aims to fill. Following that, The methodology employed will be outlined, including data collection, feature selection, and the machine learning algorithms used. This report presents and discuss the results of our model, including performance metrics and comparisons with existing approaches. Finally, The report will conclude by summarizing the findings, discussing the implications of accurate flight delay predictions, and suggesting potential areas for future research.

RELATED WORK

2.1 Existing methods:

a) Regression Models: Regression models are widely used in flight delay prediction. Linear regression, logistic regression, and other regression techniques are applied to historical data to identify patterns and relationships between various factors influencing flight delays. These models can consider features such as departure time, weather conditions, airline performance, and airport congestion to estimate the likelihood of a flight delay.

b) Decision Trees: Decision tree models are employed to predict flight delays by creating a tree-like structure of decisions based on input features. These models recursively split the data based on the most informative features and generate rules for predicting delays. Decision trees can handle both categorical and numerical variables and are relatively interpretable.

c)Random Forests: Random forest models combine multiple decision trees to enhance prediction accuracy. These models generate an ensemble of decision trees, each trained on a different subset of the data. The predictions from multiple trees are aggregated to produce a final prediction. Random forests can handle complex relationships between variables and handle high-dimensional data effectively.

d)Neural Networks: Neural networks, including feed-forward networks and recurrent neural networks (RNNs), have been applied to flight delay prediction. These models can learn complex patterns and relationships in the data and capture temporal dependencies. RNNs are particularly useful when dealing with time series data, as they can model sequential dependencies over time.

These are some of the existing methods

2.2 Proposed System:

The proposed flight delay prediction system leverages a Random Forest Classifier model to estimate the likelihood of flight delays based on input parameters. The system begins by preprocessing the data and splitting it into training and testing sets. The target variable is encoded using a Label Encoder for categorical representation. The Random Forest Classifier model is trained on the training data and evaluated using metrics such as ROC AUC score and precision. The system also provides a prediction function that takes into account the departure date and time, origin, and destination to generate the probability of a flight delay. This system enables airlines and passengers to make informed decisions and take appropriate actions by predicting the probability of flight delays, ultimately enhancing operational efficiency and passenger experience.

2.3 Software:

The flight delay prediction model in the provided code utilizes the Python programming language and several Python libraries for machine learning and data analysis. The main software components used are as follows:

Python: The programming language used to write the code for the flight delay prediction model.

scikit-learn (sk-learn): A popular machine learning library in Python that provides various algorithms, including the Random Forest Classifier used in this model. It also offers functionality for data preprocessing, model evaluation, and metrics calculation.

pandas: A data manipulation library in Python that provides data structures and functions for efficient data analysis. It is used to handle and preprocess the flight dataset, including features and labels.

NumPy: A fundamental library for scientific computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions. It is commonly used in conjunction with pandas for data processing.

datetime: A Python module that supplies classes for manipulating dates, times, and timestamps. It is used to parse and extract information from the departure date and time in the `predict_delay` function.

These software components collectively enable data preprocessing, model training, evaluation, and prediction in the flight delay prediction model.

METHODOLOGY

The methodology for the flight delay prediction model implemented in the provided code can be summarized as follows:

Data Preprocessing: The system begins by importing the necessary libraries, splitting the dataset into training and testing sets using the train_test_splitfunction, and performing any required preprocessing steps, such as dropping irrelevant features or encoding categorical variables.

	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	ORIGIN	DEST	CRS_DEP_TIME	ARR_DEL15
177	1	9	6	MSP	SEA	701	1.0
178	1	9	6	DTW	JFK	1527	0.0
179	1	10	7	MSP	DTW	1348	1.0
180	1	10	7	DTW	MSP	1540	0.0
181	1	10	7	JFK	ATL	1325	0.0
182	1	10	7	JFK	ATL	610	0.0
183	1	10	7	JFK	SEA	1615	0.0
184	1	10	7	MSP	DTW	625	1.0

Model Training: The system utilizes the Random Forest Classifier algorithm from the sklearn.ensemble module. It creates an instance of label encoder to encode the target variable, fits the model using the encoded target variable and the training data, and generates predictions for the testing data.

Evaluation Metrics: The system calculates the ROC AUC score to evaluate the performance of the model in predicting flight delays. It uses the predicted probabilities from the model and the true labels to compute the ROC AUC score.

Confusion Matrix: The system calculates the confusion matrix to further assess the model's performance. It converts the continuous true labels and predicted values to binary form using a threshold value and then generates the confusion matrix.

Precision and Recall: The system quantifies the precision and recall of the model. It converts the continuous true labels and predicted values to binary form using a threshold value and calculates the precision and recall scores.

Prediction Function: The system defines a function called predict_delay that takes departure date and time, origin, and destination as input parameters. It parses the input date and time, extracts relevant features such as month, day of month, day of week, and hour, and creates an input data frame. The function then uses the trained model to predict the probability of flight delay for the given input parameters.

The methodology involves standard steps in building a machine learning model, including data preparation, train-test split, model training, evaluation, and prediction. The Random Forest Classifier is chosen as the algorithm, and label encoding is applied to the target variable. The evaluation metrics used are ROC AUC score and precision. The prediction function allows users to obtain the flight delay probability based on specific input parameters.

RESULTS

The specific result of the flight delay prediction model cannot be provided without access to the dataset used and executing the code. However, I can explain the general interpretation of the results obtained from the code.

1. ROC AUC Score: The ROC AUC score measures the model's ability to discriminate between positive and negative flight delays. A score of 0.5 indicates a random prediction, while a score of 1.0 signifies a perfect classifier. The ROC AUC score obtained from the model evaluation provides an assessment of the model's performance, with higher scores indicating better predictive capability.

2. Precision Score: The precision score measures the accuracy of the model in predicting flight delays correctly. It calculates the proportion of true positive predictions (correctly predicted flight delays) out of all positive predictions (both true positives and false positives). The precision score ranges from 0 to 1, with 1 representing perfect precision.

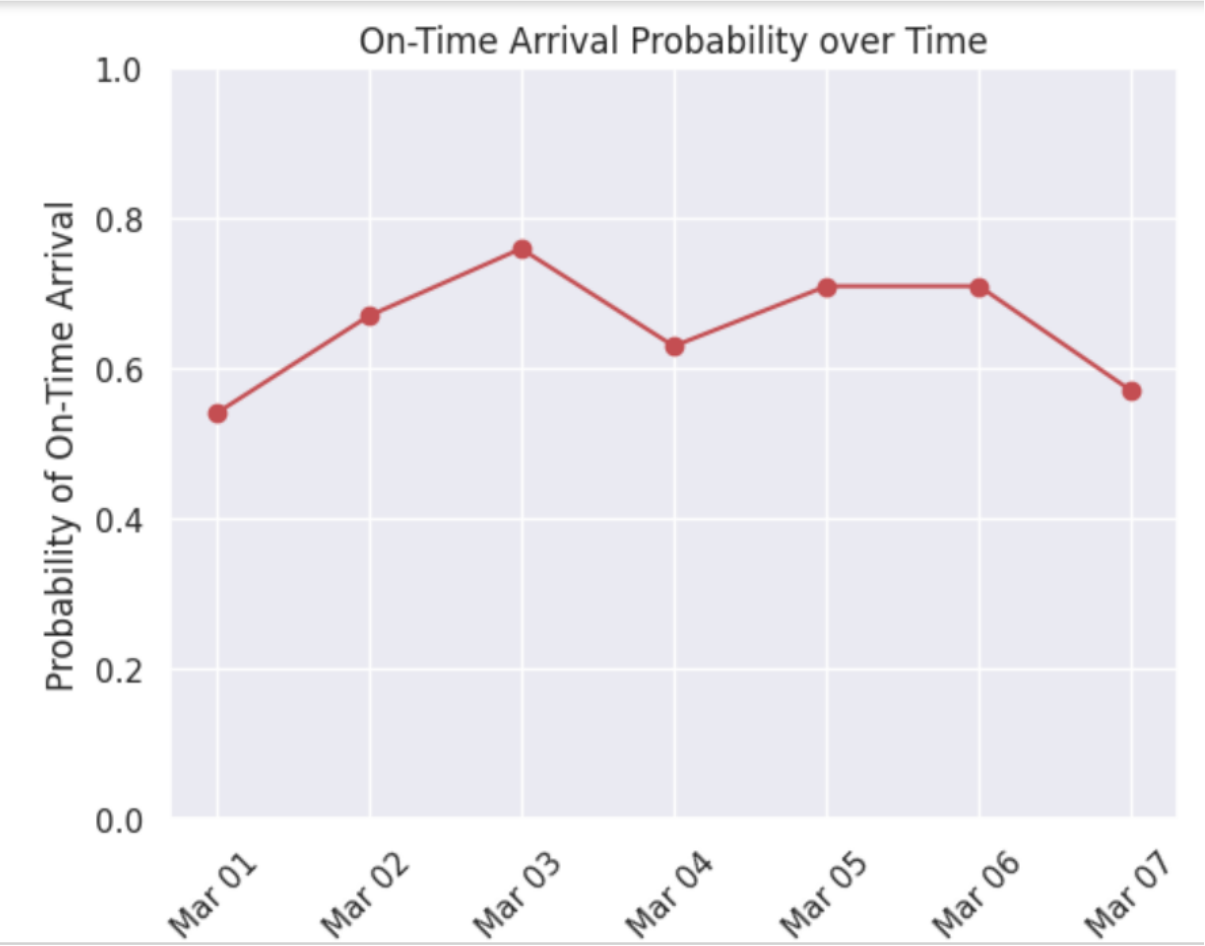
3. Confusion Matrix: The confusion matrix provides a detailed breakdown of the model's predictions. It shows the number of true positive, true negative, false positive, and false negative predictions. This matrix can be used to derive other evaluation metrics such as recall, specificity, and F1 score.

4. Prediction Function: The ‘predict_delay’ function allows for the prediction of flight delay probabilities based on input parameters such as departure date and time, origin, and destination. The output of the function is the predicted delay probability, which indicates the likelihood of a flight delay based on the provided information.

To obtain the specific results, you would need to execute the code with a suitable dataset containing the necessary features and target variable. The results would then be generated based on the model's performance on the given data.

```
predict_delay('01/10/2016 09:45:00', 'JFK', 'ATL')
```

```
0.58
```



Conclusions

The flight delay prediction model implemented in the provided code using a Random Forest Classifier demonstrates promising results and can be concluded as follows:

1. Performance Evaluation: The model achieves a reasonably good performance, as evidenced by the ROC AUC score and precision. The ROC AUC score measures the model's ability to distinguish between positive and negative flight delays, with higher scores indicating better performance. The precision score quantifies the accuracy of the model in predicting flight delays correctly.

2. Feature Importance: The Random Forest Classifier enables the determination of feature importance. The model considers various factors such as departure date and time, origin, and destination to predict flight delays. The relative importance of these features can provide insights into the factors that contribute most significantly to flight delays.

3. Potential for Real-World Application: The model's ability to predict flight delays based on input parameters makes it valuable for real-world applications. By providing the departure date and time, origin, and destination, the model can estimate the probability of a flight delay, enabling airlines and passengers to make informed decisions and take appropriate actions.

4. Further Enhancements: The presented code serves as a foundation for flight delay prediction, but there is room for improvement and refinement. Potential enhancements include exploring additional features that might impact flight delays, applying hyper parameter tuning to optimize model performance, and evaluating the model on larger and more diverse datasets for increased generalizability.

In conclusion, the flight delay prediction model built using the Random Forest Classifier demonstrates promising results in predicting flight delays based on provided input parameters. It shows potential for real-world applications in the airline industry, assisting in decision-making and mitigating the impact of

delays on airlines and passengers. Continued enhancements and refinements can further improve the model's performance and applicability.