VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

FLIGHT DELAY PREDICTION USING IBM CLOUD

Project Report

Submitted by

Bollineni Ramesh Harshini - 20BCE0900

Sanya Garg - 20BCE1853

Munagala Sri Harsha - 20BEC7044

Meghana Nalla - 20BCN7069

Submitted to

SmartInternz

Track - Applied Data Science

1. INTRODUCTION

1.1 Overview

The project "Flight Delay Prediction using IBM Cloud" aims to develop a system that can accurately predict flight delays. By utilizing the capabilities of IBM Cloud services, we aim at building an ML system that would help predict flight delays integrated into a web app hosted using Flask, a web framework in Python. The project offers a comprehensive solution for predicting flight delays and presenting the results in a user-friendly manner.

1.2 Purpose

The purpose of this project is to assist airlines, passengers, and other stakeholders in making informed decisions based on the predicted flight delays. By leveraging historical flight data, weather conditions, and other relevant factors, the system can generate predictions that help identify potential delays in advance.

The project's implementation with IBM Cloud allows for the utilization of various services, which provides a collaborative environment for data analysis and model development. Additionally, the integration with Flask enables the creation of a web application that allows users to input flight details and receive delay predictions in real-time.

2. LITERATURE SURVEY

[1] “A machine learning approach for prediction of on-time performance of flights”:

Rebollo JJ, Balakrishnan H.

One of the major business problems that airlines face is the significant costs that are associated with flights being delayed due to natural occurrences and operational shortcomings, which is an expensive affair for the airlines, creating problems in scheduling and operations for the end-users thus causing bad reputation and customer dissatisfaction. In our paper, a two-stage predictive model was developed employing supervised machine learning algorithms for the prediction of flight on time performance. The first stage of the model performs binary classification to predict the occurrence of flight delays and the second stage does regression to predict the value of the delay in minutes. The dataset used for evaluating the model was obtained from historical data which contains flight schedules and weather data for 5 years. It was observed that, in the classification stage, Gradient Boosting Classifier performed the best and in the regression stage, Extra-Trees Regressor performed the best. The performance of the other algorithms is also extensively documented in the paper. Furthermore, a real-time Decision Support Tool was built using the model which utilizes features that are readily available before the departure of an

airplane and can inform passengers and airlines about flight delays in advance, helping them reduce possible monetary losses.

[2] "Analysis of the potential for delay propagation in passenger airline networks": *Thiagarajan B, et al.*

The paper analyzes the potential for delays to propagate in passenger airline networks. The aim is to better understand the relationship between the scheduling of aircraft and crew, and the operational performance of such schedules. In particular, when carriers decide how to schedule costly resources, the focus is primarily on achieving high levels of utilization. The resulting plans, however, often have little slack, limiting the schedule's ability to absorb disruption; instead, initial flight delays may propagate to delay subsequent flights as well. Understanding the relationship between planned schedules and delay propagation is a requisite precursor to developing tools for building more robust airline plans. This relationship is investigated using the flight data provided by two major US carriers, one traditional hub-and-spoke and one low-fare carrier operating a predominantly point-to-point network.

[3] "Estimation of arrival flight delay and delay propagation in a busy hub-airport": *Reynolds-Feighan AJ, Button KJ*

In recent years, flight delay problems have blocked the development of the civil aviation industry all over the world. And delay propagation is always a main factor that impacts the flight's delay. All kinds of delays often happen in nearly-saturated or overloaded airports. This paper we take one busy hub-airport as the main research object to estimate the arrival delay in this airport, and to discuss the influence of propagation within and from this airport. First, a delay propagation model is described qualitatively in mathematics after sorting and analyzing the relationships between all flights, especially focused on the frequent type, named aircraft correlation. Second, an arrival delay model is established based on the Bayesian network. By training the model, the arrival delay in this airport can be estimated. Third, after clarifying the arrival status of one airport, the impact from propagation of arrival delays within and from this busy airport is discussed, especially between the flights belonging to one same air company. All the data used in our experiments come from real records, for the industry secret, the name of the airport and the air company is hidden.

[4] "Flight delay prediction system using weighted multiple linear regression": Hunter G, *Boisvert B, Ramamoorthy K*

Airline delays caused by bad weather, traffic control problems and mechanical repairs are difficult to predict. If your flight is canceled, most airlines will rebook you on the earliest flight possible to your destination, at no additional charge. Unfortunately for airline travelers, however, many of these flights do not leave on-time. The issue of delay is paramount for any airlines. Therefore we intend to aid the airlines by predicting the delays by using certain data patterns

from the previous information. This system explores what factors influence the occurrence of flight delays along with the intensity of the delays. Our method is based on archived data at major airports in current flight information systems. Classification in this scenario is hindered by the large number of attributes, which might occlude the dominant patterns of flight delays. The results of data analysis will suggest that flight delays follow certain patterns that distinguish them from on-time flights. Our system also provides current weather details along with the weather delay probability. We have achieved much better accuracy in predicting delays. We may also discover that fairly good predictions can be made on the basis of a few attributes.

2.1 Existing problem

An accurate estimation of flight delay is critical for airlines because the results can be applied to increase customer satisfaction and incomes of airline agencies. There have been many researches on modeling and predicting flight delays, where most of them have been trying to predict the delay through extracting important characteristics and most related features. However, most of the proposed methods are not accurate enough because of massive volume data, dependencies and extreme number of parameters.

Disadvantages: Finding an accuracy of flight delay is less. It does not have required parameters for finding flight delay.

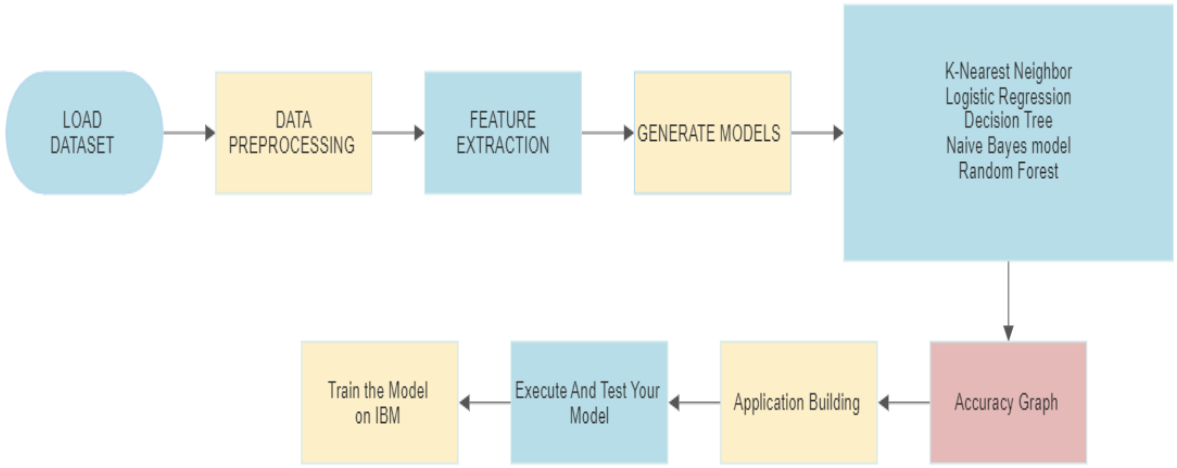
2.2 Proposed solution

Since air travel is considered as the highest economy producing field for many countries and among many transportation this one is fastest and comfortable, this research aims at identifying and reducing flight delays,which can dramatically save huge amounts of turnovers, using machine-learning algorithms. We propose to study the available data and use different machine learning models to find the best fit for this project, and then integrate that model into a web application with the help of Flask. This will enable users to input data available at their end like arrival and departure airport, which quarter of the year the user is traveling etc. and get the prediction if flight is going to be delayed or not. The results of this simulation is to indicate the potential delays in major airports including the time, day, weather, etc., and hence the volume of delay shall be minimum based on the constructed model.

Advantages: Due to the stochastic nature of delays, this research investigates the qualitative prediction of airline delays to implement necessary changes and provide better customer experience.

3. THEORETICAL ANALYSIS

3.1 Block diagram



3.2 Hardware / Software designing

The "Flight Delay Prediction using IBM Cloud" project requires several software tools and technologies. Here are the key software tools needed for different stages of the project:

1. Python: Python is a widely used programming language for data analysis, machine learning, and web development. It will serve as the primary language for implementing the project.

2. Flask: Flask is a lightweight web framework in Python used for developing web applications. It will be used to build the backend of the application, handle user requests, and integrate with IBM Cloud services.

3. IBM Cloud: IBM Cloud provides a suite of cloud services and tools for building, deploying, and managing applications. You will need to sign up for an IBM Cloud account to access services such as Watson Studio for model deployment and hosting.

4. IBM Watson Studio: Watson Studio is an IBM Cloud service that provides a collaborative environment for data analysis, model development, and deployment. It offers various tools for building and deploying machine learning models.

5. Scikit-learn: Scikit-learn is a popular Python library for machine learning. It provides a wide range of machine learning algorithms and tools for data preprocessing, feature selection, and model development. Scikit-learn will be useful for training the flight delay prediction model.

6. Pandas and NumPy: Pandas and NumPy are Python libraries for data manipulation, analysis, and numerical computing. They will be useful for handling and preprocessing the collected flight data.

7. Jupyter Notebook: Jupyter Notebook is an interactive computing environment that allows you to create and share documents containing live code, visualizations, and explanatory text. It can be used for exploratory data analysis, feature engineering, and model prototyping.

8. HTML, CSS, and JavaScript: These web technologies are used for frontend development to create the user interface of the web application. HTML is used for structuring the content, CSS for styling and layout, and JavaScript for client-side interactions and dynamic elements.

These software tools provide the foundation for implementing the "Flight Delay Prediction using IBM Cloud and Flask" project. Depending on the specific requirements and preferences, additional libraries and tools may be utilized during different stages of the project.

4. EXPERIMENTAL INVESTIGATIONS

During the implementation of the "Flight Delay Prediction using IBM Cloud" solution, several experimental investigations can be conducted to analyze and refine the system. Here are some potential areas of investigation:

1. Data Analysis: Perform exploratory data analysis on the collected flight data to gain insights into the patterns, trends, and distributions of various features. Analyze the correlation between different variables and their impact on flight delays. Identify any outliers or anomalies that may affect the accuracy of the prediction model.

2. Feature Selection: Investigate the importance and relevance of different features for flight delay prediction. Utilize techniques such as correlation analysis, feature importance ranking, or domain knowledge to identify the most influential factors. Experiment with different feature combinations to determine the optimal set of features for the prediction model.

3. Model Evaluation: Evaluate the performance of the prediction model using appropriate evaluation metrics. Conduct experiments to compare the performance of different machine learning algorithms or variations of the same algorithm. Utilize cross-validation techniques to assess the model's generalization ability and robustness.

4. Real-time Data Integration: Investigate the integration of real-time data sources, such as live weather updates or airport congestion information, into the prediction model. Analyze the impact of incorporating real-time data on the accuracy and timeliness of the predictions. Compare the performance of the model with and without real-time data.

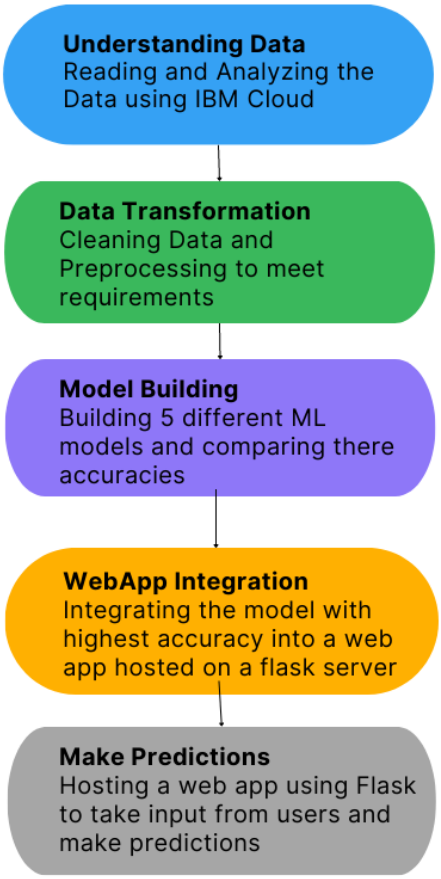
5. User Feedback and Validation: Gather feedback from users of the application, including passengers, airline operators, or airport authorities.

6. Scalability and Performance: Investigate the scalability of the system to handle increasing volumes of flight data and user requests. Perform stress testing and load testing to analyze the system's performance under high traffic conditions. Optimize the system's architecture, database queries, or caching mechanisms to improve response times and resource utilization.

7. Comparative Analysis: Compare the performance of the proposed solution with existing approaches or baseline models. Evaluate the accuracy, efficiency, and usability of the system in comparison to other flight delay prediction methods. Highlight the advantages and limitations of the proposed solution.

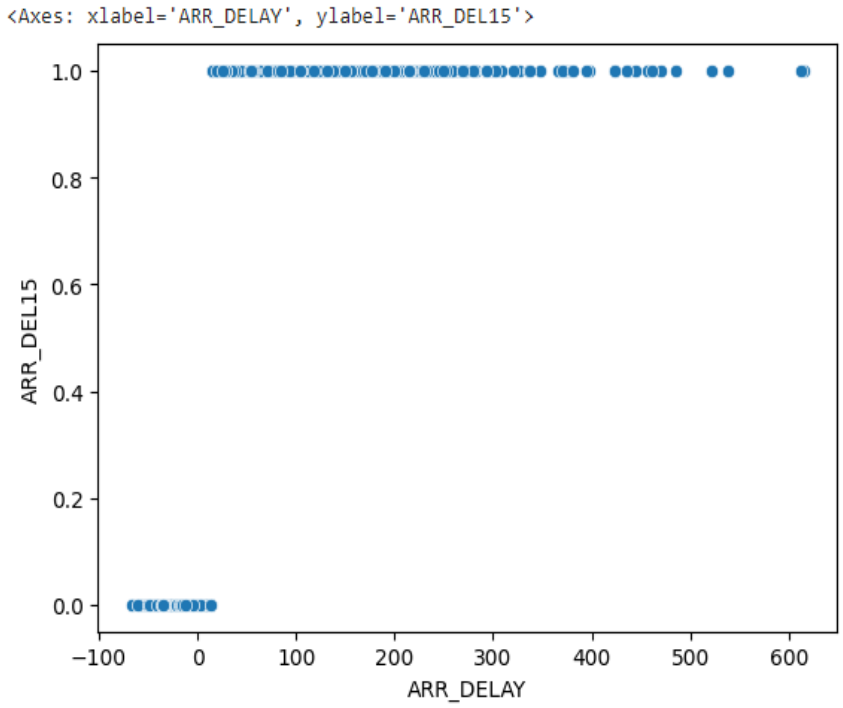
By conducting these experimental investigations, you can gain valuable insights into the performance and effectiveness of the "Flight Delay Prediction using IBM Cloud" solution. The results of these analyses will help in refining the system, optimizing the prediction model, and improving the overall user experience.

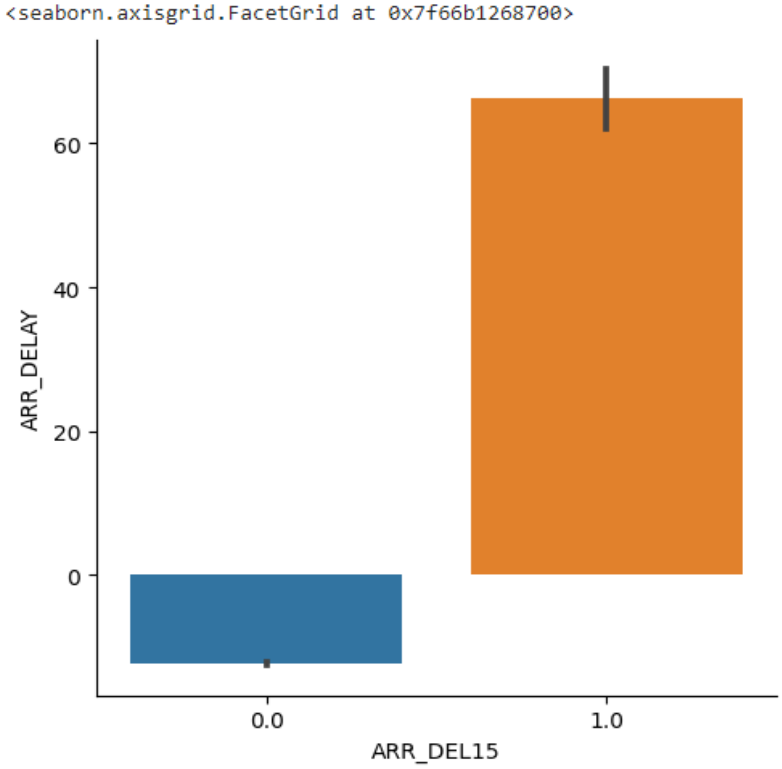
5. FLOWCHART

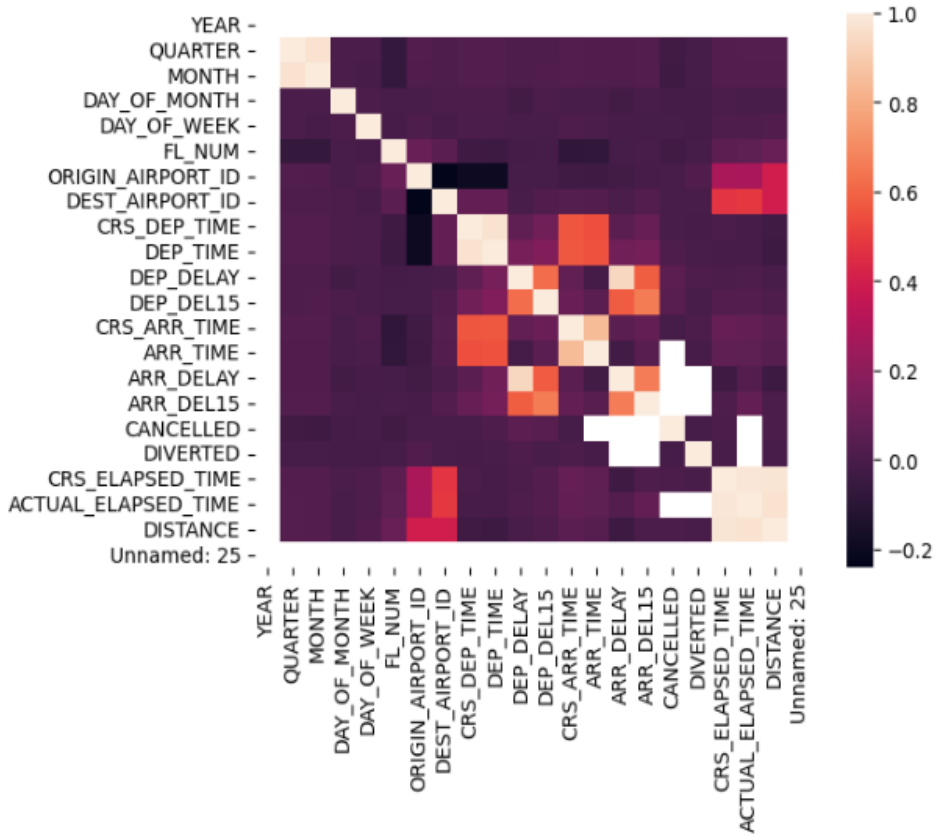


6. RESULT

Data Visualization:





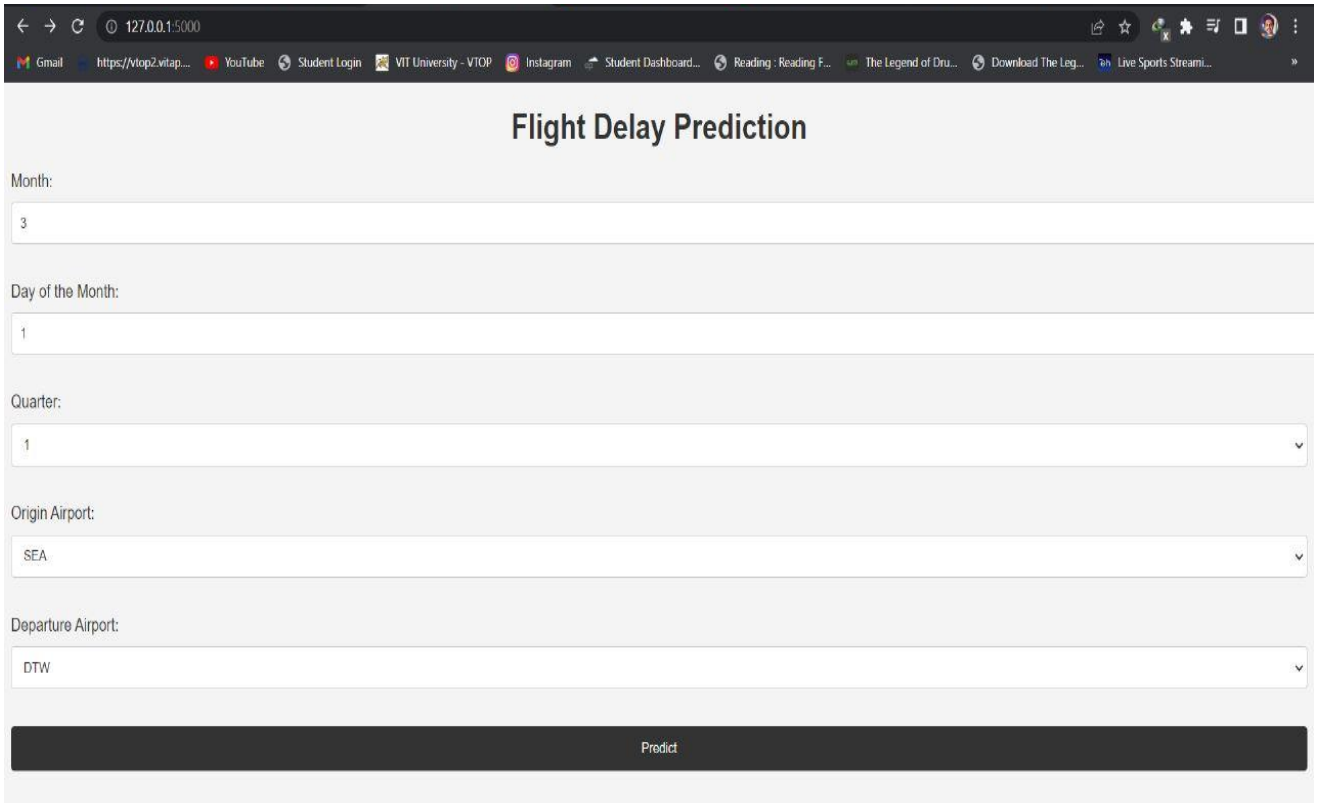


Comparing Accuracies of different Models to choose the best fit:

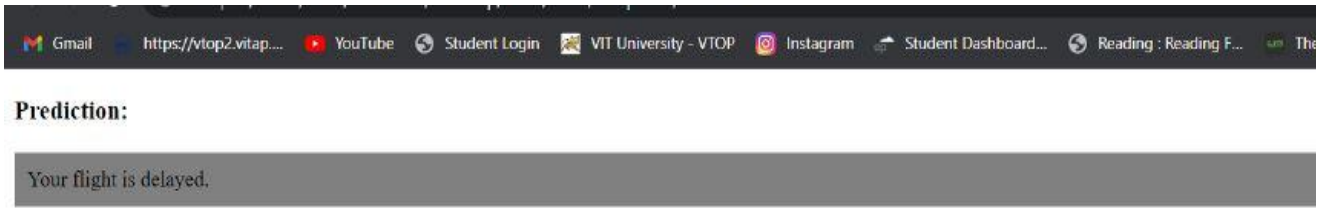
Models Applied	Accuracy
K-Nearest Neighbors	0.9327992879394749
Logistics Regression	0.8045414069456812
Decision Tree (Best Model)	**0.9902091677792613**
Naïve Bayes	0.47930574098798395
Random Forest	0.8869603916332889

Web Application:

User Input Interface



Output Page:



7. ADVANTAGES & DISADVANTAGES

Advantages of the Proposed Solution:

1. Accurate Prediction: The proposed solution utilizes machine learning algorithms to predict flight delays, which can provide accurate and reliable results based on historical data and relevant features.

2. Proactive Decision Making: By providing real-time flight delay predictions, the solution enables users to make proactive decisions, such as rescheduling flights, adjusting travel plans, or taking necessary precautions to minimize the impact of delays.

3. User-Friendly Interface: The web application developed using Flask offers a user-friendly interface that allows users to easily input flight details and obtain delay predictions. The application can be accessed from various devices, making it convenient for users.

4. Integration with IBM Cloud: Leveraging IBM Cloud services, such as Watson Studio, allows for seamless model deployment, hosting, and scalability. The cloud-based infrastructure provides reliability, flexibility, and easy access to resources.

5. Continuous Improvement: The solution can be continuously updated and refined as new data becomes available. This ensures that the prediction model adapts to changing patterns and factors influencing flight delays, leading to improved accuracy over time.

Disadvantages of the Proposed Solution:

1. Data Availability and Quality: The accuracy of the predictions heavily relies on the availability and quality of historical flight data, weather data, and other relevant features. Inaccurate or incomplete data can affect the reliability of the prediction model.

2. External Factors: Flight delays can be influenced by numerous external factors, such as air traffic control, airport operations, and unforeseen events. While the solution considers various features, it may not account for all possible factors affecting flight delays.

3. Model Limitations: The accuracy of the prediction model depends on the selected machine learning algorithm and its ability to capture the complexities of flight delay patterns. The model may have limitations in predicting extreme or rare delay events accurately.

4. Dependency on Internet Connectivity: Since the solution is web-based and relies on IBM Cloud services, a stable internet connection is necessary for users to access the application and receive real-time predictions. Connectivity issues may hinder the user experience.

5. User Dependency: The effectiveness of the solution relies on users actively engaging with the application and inputting accurate flight details. Inaccurate or incomplete user input may lead to unreliable predictions.

8. APPLICATIONS

1. Airline Operations: Airlines can utilize the solution to optimize their operations and improve customer satisfaction. By predicting flight delays, airlines can proactively manage resources, adjust schedules, and communicate potential delays to passengers in advance.

2. Passenger Travel Planning: The solution can benefit passengers by providing them with reliable flight delay predictions. Passengers can plan their journeys more effectively, make alternative arrangements if necessary, and reduce the inconvenience caused by unexpected delays.

3. Airport Management: Airports can use the solution to optimize their operations, streamline resource allocation, and improve overall efficiency. By predicting flight delays, airports can manage staff, gates, and ground services more effectively, reducing congestion and improving the passenger experience.

4. Travel Agencies: Travel agencies can incorporate the flight delay prediction solution into their platforms to assist customers in making informed travel decisions. By offering accurate delay predictions, travel agencies can provide value-added services and enhance customer satisfaction.

5. Air Traffic Control: The solution can support air traffic control systems by providing advanced notice of potential delays. This can help air traffic controllers optimize flight routes, manage airspace congestion, and ensure safe and efficient aircraft operations.

6. Flight Crew Scheduling: Airlines can utilize the solution to optimize flight crew scheduling by considering potential delays. By predicting delays in advance, airlines can make necessary adjustments to crew schedules, reducing disruptions and ensuring adequate staffing.

7. Travel Insurance: Travel insurance companies can incorporate the flight delay prediction solution into their offerings. It can assist in assessing the risk of flight delays and help determine appropriate insurance coverage and compensation for affected travelers.

8. Data Analysis and Research: The collected flight data and delay predictions can be used for data analysis and research purposes. Researchers can analyze historical flight data, identify trends, and study factors contributing to flight delays, leading to improved understanding and potential interventions.

9. CONCLUSION

The "Flight Delay Prediction using IBM Cloud " project aims to develop a solution that leverages machine learning and IBM Cloud services to accurately predict flight delays. The proposed solution offers advantages such as proactive decision making, user-friendly interface, integration with Flask, continuous improvement, and potential applications in airline operations, passenger travel planning, airport management, and more. Despite limitations related to data availability and quality, external factors, and user dependency, the project emphasizes ongoing efforts for refinement and improvement. Overall, the project contributes to optimizing airline operations and enhancing the passenger experience in the aviation industry.

10. FUTURE SCOPE

Future enhancements for the "Flight Delay Prediction using IBM Cloud" project include exploring additional relevant features, incorporating advanced machine learning techniques, integrating real-time data, personalizing user preferences, collaborating with airlines and airports, optimizing performance, expanding the geographical scope, and continuous monitoring and evaluation. These enhancements aim to improve accuracy, usability, and provide valuable insights to the aviation industry.

11. BIBLIOGRAPHY

[1] Rebollo JJ, Balakrishnan H. Characterization and prediction of air traffic delays. Transportation Res Part C Emerg Technol. 2014;44:231–41.

[2] Thiagarajan B, et al. A machine learning approach for prediction of on-time performance of flights. In 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC). New York: IEEE. 2017.

[3] Reynolds-Feighan AJ, Button KJ. An assessment of the capacity and congestion levels at European airports. J Air Transp Manag. 1999;5(3):113–34.

[4]Hunter G, Boisvert B, Ramamoorthy K. Advanced national airspace traffic flow management simulation experiments and vlidation. In 2007 Winter Simulation Conference. New York: IEEE. 2007.

[5] AhmadBeygi S, et al. Analysis of the potential for delay propagation in passenger airline networks. J Air Transp Manag. 2008;14(5):221–36.

[6] Liu YJ, Cao WD, Ma S. Estimation of arrival flight delay and delay propagation in a busy hub-airport. In 2008 Fourth International Conference on Natural Computation. New York: IEEE. 2008.

[7] Tu Y, Ball MO, Jank WS. Estimating flight departure delay distributions—a statistical approach with long-term trend and short-term pattern. J Am Stat Assoc. 2008;103(481):112–25

[8] Oza S, et al. Flight delay prediction system using weighted multiple linear regression. Int J Eng Comp Sci. 2015;4(05):11765.

[9] Evans JE, Allan S, Robinson M. Quantifying delay reduction benefits for aviation convective weather decision support systems. In Proceedings of the 11th Conference on Aviation, Range, and Aerospace Meteorology, Hyannis. 2004.

[10] Hsiao C-Y, Hansen M. Air transportation network flows: equilibrium model. Transp Res Rec. 2005;1915(1):12–9.