PROJECT REPORT

PROJECT TITLE : DEVELOPING A FLIGHT DELAY PREDICTION MODEL USING MACHINE LEARNING

TEAM ID : PNT2022TMID37576

TEAM MEMBERS : PREETHI S(TEAM LEAD)
MADHUBALA A
ROHINI S
AKSHAYA S

INTRODUCTION

1. Project overview

In the past two decades, air travel has become increasingly popular and has become increasingly accessible to people all over the world. Aviation has evolved to become one of the most important forms of transportation, with its efficiency and reliability making it the preferred choice for long-distance travel. However, flight delays are a major problem in the aviation industry, and they are becoming more and more common. In the United States, the average delay has increased by 30% since 2000, and the cost of delays has risen to $32 billion per year. There are many factors that can contribute to flight delays, such as weather, air traffic control, and maintenance. However, the most common cause of delays is simply that the plane is not ready to take off on time. This is usually due to the fact that the plane is not fully loaded with passengers, baggage, and fuel. It can also be due to technical problems with the plane itself. The goal of this project is to develop a machine learning model that can predict flight delays. The model will be trained on a dataset of historical flight data, and it will be used to predict the delay of a flight before it even takes off.

2. Purpose

The purpose of this project is to develop a machine learning model that can predict flight delays. The model will be trained on a dataset of flight information, and will be used to predict the arrival delay of flights. The project is divided into two parts:

1. Data pre-processing and feature engineering
2. Model training and testing

In the first part, the data will be pre-processed and features will be engineered. This part will be focused on cleaning the data and making sure that the features are suitable for training the machine learning model. In the second part, the machine learning model will be trained and tested. This part will focus on tuning the model to get the best performance possible.

LITERATURE SURVEY

1. Existing problem

Airlines, airports, and passengers would all benefit from a more accurate flight delay prediction model. Currently, models used by airlines to predict flight delays are based on historical data and do not take into account real-time data such as weather conditions. This can lead to delays and cancellations, as well as increased costs for airlines.

2. References

1. Khaksar, H., & Sheikholeslami, A. (2017). Airline delay prediction by machine learning algorithms. *Scientia*
2. Esmaeilzadeh, E., & Mokhtarimousavi, S. (2020). Machine learning approach for flight departure delay prediction and analysis. *Transportation*

Research Record: Journal of the Transportation Research Board, *2674*(8), 145–159.

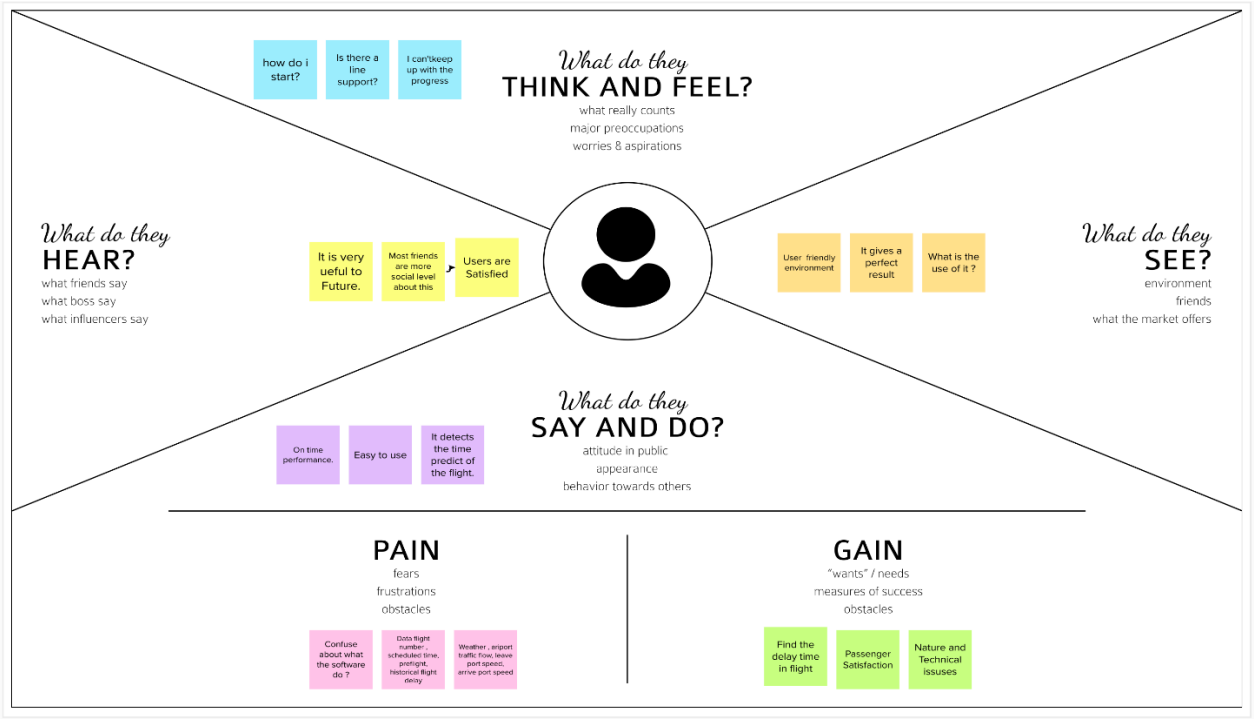
[3] M. Al-Tabbakh, S., M. Mohamed, H., & H. El, Z. (2018). Machine learning techniques for analysis of Egyptian flight delay. *International Journal of Data Mining & Knowledge Management Process*, *8*(3), 01–14’

4. Ye, B., Liu, B., Tian, Y., & Wan, L. (2020). A methodology for predicting aggregate flight departure delays in airports based on supervised learning. *Sustainability*, *12*(7), 2749.
5. ATLIOĞLU, M. C., BOLAT, M., ŞAHİN, M., TUNALI, V., & KILINÇ, D. (2020). Supervised learning approaches to flight delay prediction. *Sakarya University Journal of Science*.
6. Yu, B., Guo, Z., Asian, S., Wang, H., & Chen, G. (2019). Flight delay prediction for commercial air transport: A deep learning approach. *Transportation Research Part E: Logistics and Transportation Review*, *125*, 203–221.

Problem Statement Definition

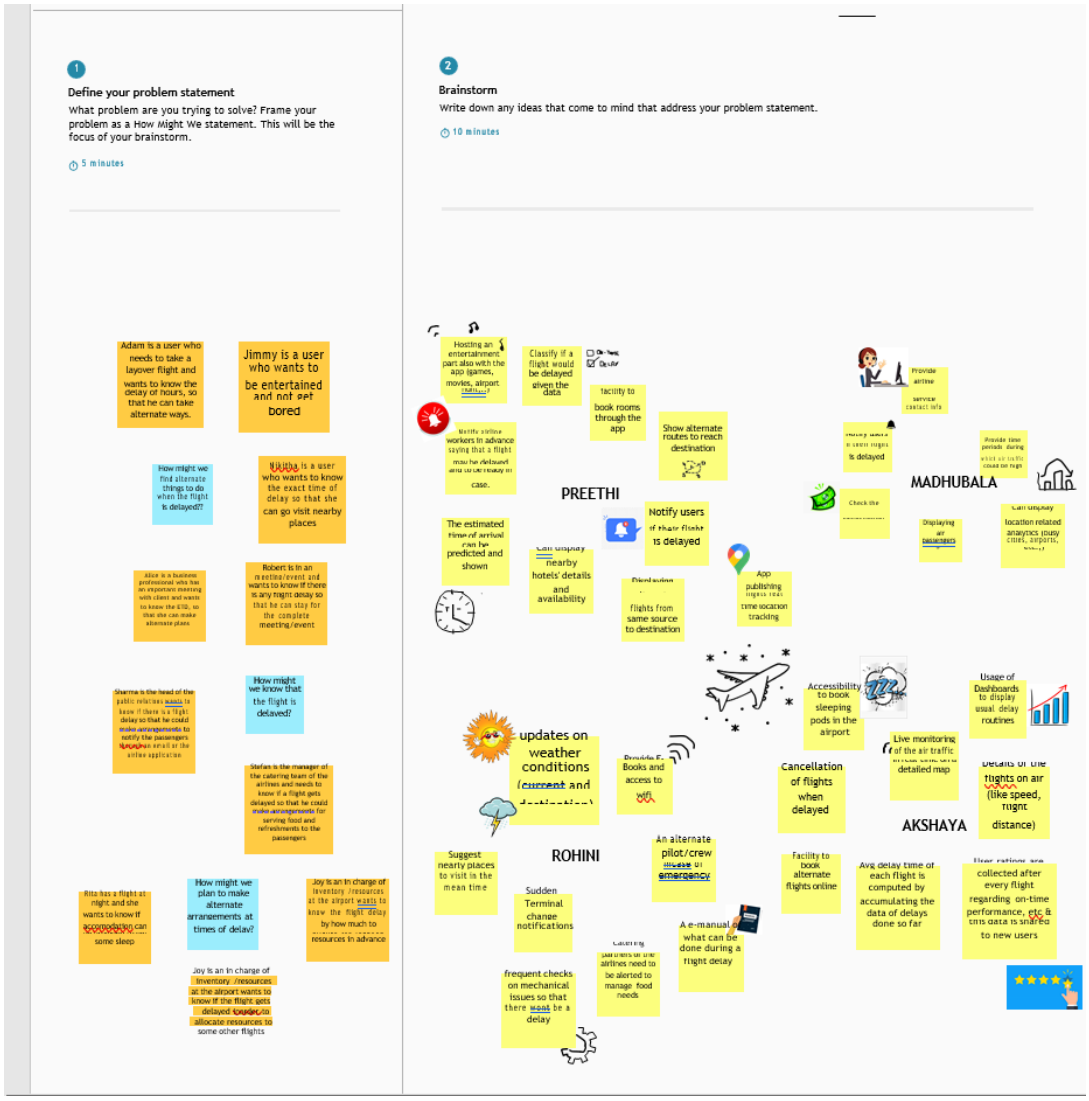
The main objective of the model is to predict flight delays accurately in order to optimize flight operations and minimize delays. These delays are responsible for large economic and environmental losses.

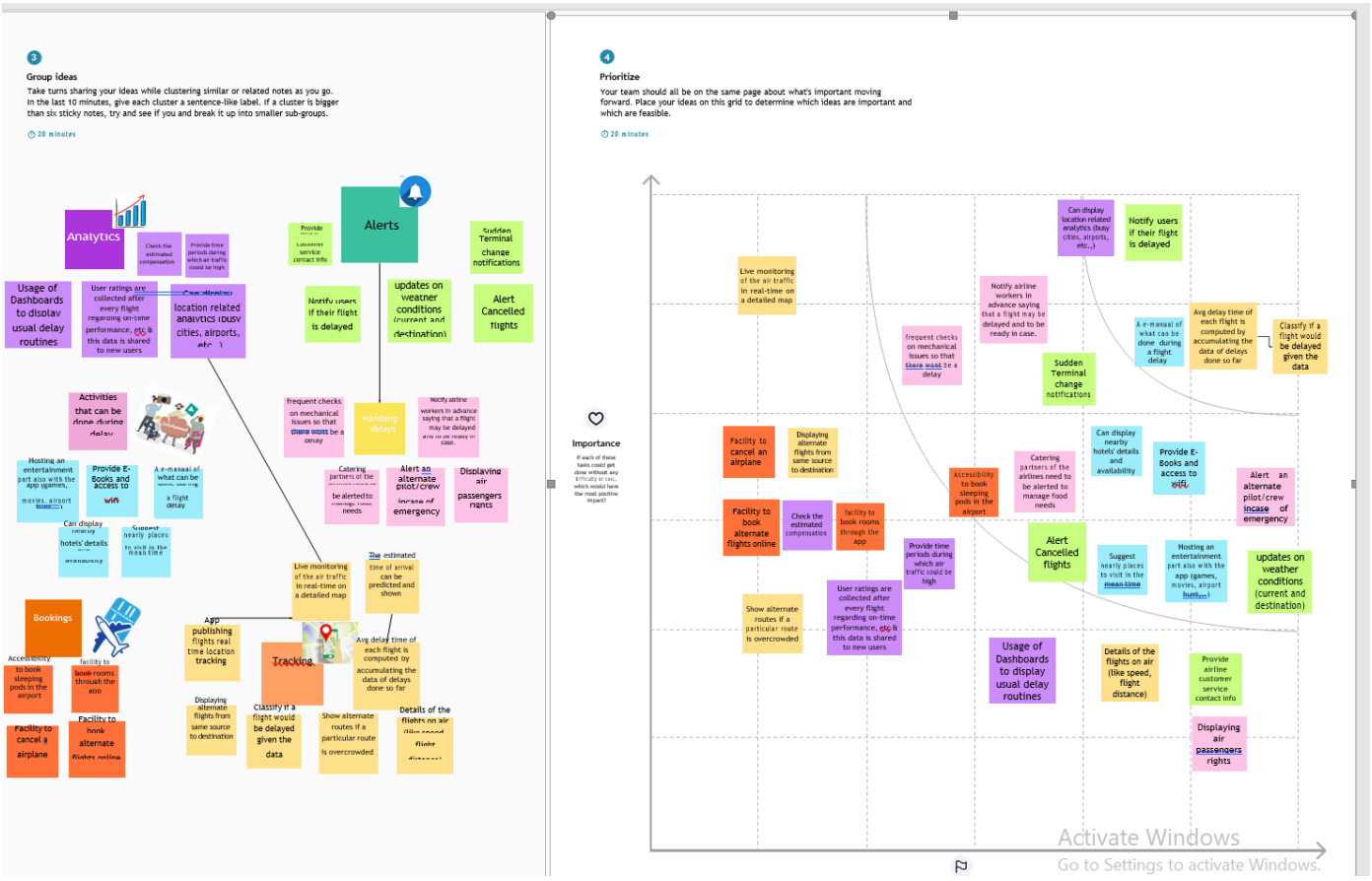
IDEATIOIN AND PROPOSED SOLUTION



Empathy Map Canvas

Ideation and Brainstroming

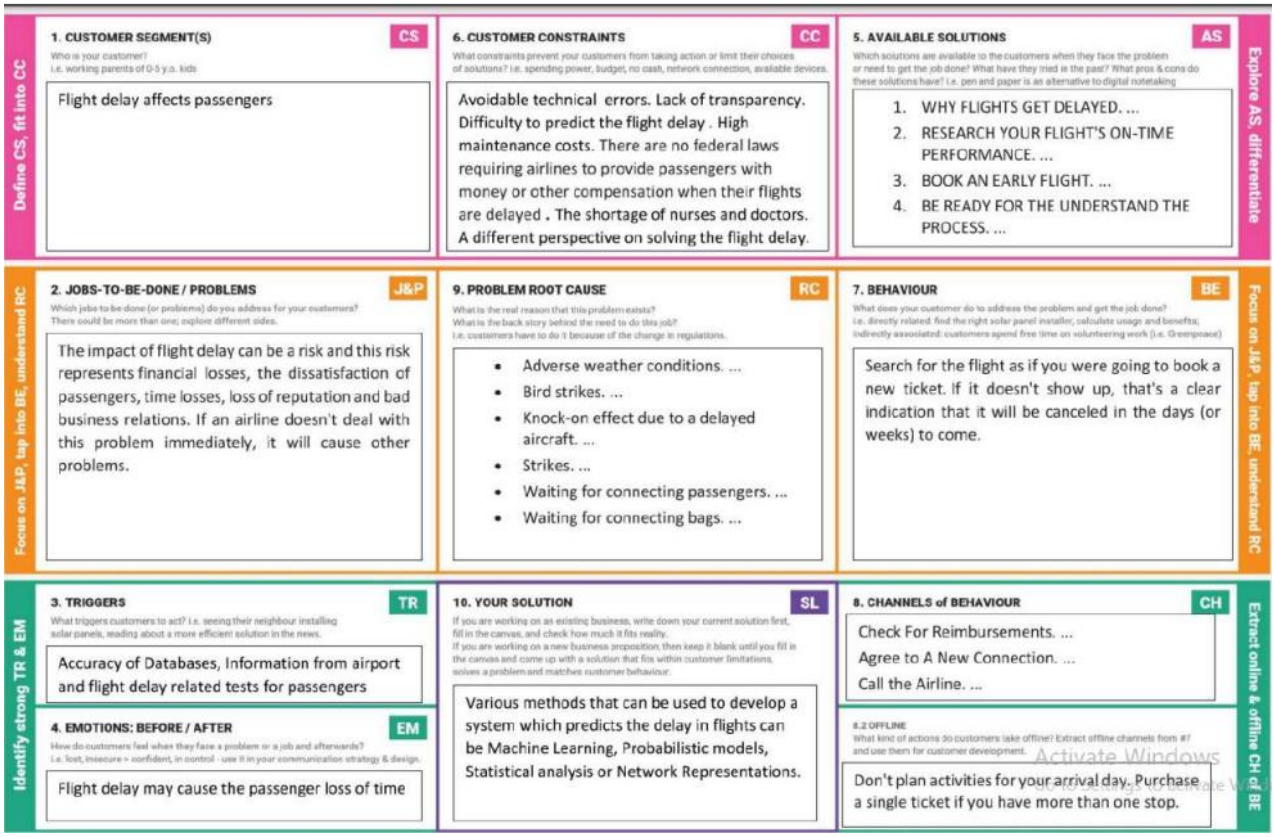



3
Group ideas
Take turns sharing your ideas while clustering similar or related notes as you go. In the last 10 minutes, give each cluster a sentence-like label. If a cluster is bigger than six sticky notes, try and see if you and break it up into smaller sub-groups.
20 minutes
Analytics
Alerts
Usage of Dashboards to display usual delay routines
Alert Cancelled flights
updates on weather conditions (current and destination)
Notify users if their flight is delayed
Sudden Terminal change notifications
Activities that can be done during delay
Handling delays
Bookings
Tracking
Live monitoring of the air traffic in real-time on a detailed map
Facility to cancel an airplane
Facility to book alternate flights online
Displaying air passengers rights
4
Prioritize
Your team should all be on the same page about what's important moving forward. Place your ideas on this grid to determine which ideas are important and which are feasible.
20 minutes
Importance
Notify users if their flight is delayed
Sudden Terminal change notifications
Alert Cancelled flights
Usage of Dashboards to display usual delay routines
Live monitoring of the air traffic in real-time on a detailed map
Facility to cancel an airplane
Facility to book alternate flights online
Alert an alternate pilot/crew incase of emergency
updates on weather conditions (current and destination)
Activate Windows
Go to Settings to activate Windows.

Proposed Solution

S.No	Parameter	Description
1.	Problem Statement	The main objective of the model is to predict flight delays accurately in order to optimize flight operations and minimize delays. These delays are responsible for large economic and environmental losses.
2.	Idea/Solution description	Using a machine learning model, we can predict flight arrival delays. A flight is considered to be delayed when the difference between scheduled and actual arrival times is greater than 15 minutes.
3.	Novelty/Uniqueness	A user friendly app that provides accurate predictions of the delay time which can be easel accessible.
4.	Social Impact/Customer Satisfaction	Predicting flight delays can improve airline operations and passenger satisfaction, which will result in a positive impact on the economy.
5.	Business Model (Revenue Model)	Make revenue from commercial advertisements and sell the model to airline companies.
6.	Scalability of the solution	This model can handle any number of inputs and provide the respective outputs.

Problem Solution Fit



REQUIREMENT ANALYSIS

1. Functional requirement

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User registration & login	Registration &login of passengers via Google with email id and password
FR-2	Detailed arrival and departure time of flights	With the flight no and name, the passenger cam see thedetails (time, boarding station, etc)of his/her in the dashboard.

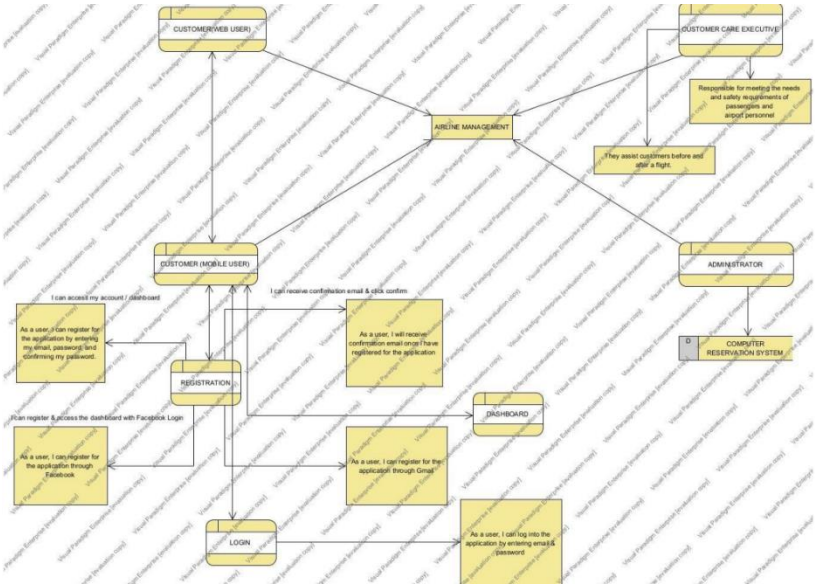
FR-3	Intimate the accurate flight timings to passengers	With the help of various machine learning algorithms, when given the right input features (actual arrival time & departure time, scheduled time, etc) we can predictthe delay in time of the flight which will also be shown in the dashboard and updated time-to-time.
FR-4	Airline helpdesk provide alternatives	The contact details of different airlines will be provided,The passenger will also be able to look for any alternative flight in case the flights get cancelled.
FR-5	Passenger feedback	The feedback will be got from the users or how theapplication was to use, with their feedback and suggestions, we can improve the application further.

2. Non-Functional requirement

FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	The application will have an easy-to-use GUI. Users will find it simple to comprehend and utilize all the capabilities of the application.
NFR-2	**Security**	The technique known as database replication will be utilised for the application security to ensure the safety of all crucial data
NFR-3	**Reliability**	The application will be consistent in all scenarios and work without fail in any environment
NFR-4	**Performance**	The applications response time is direct &faster which is determined by the efficiency of the implemented machine algorithm.
NFR-5	**Availability**	The application will be accessible to users 24 hours a day,7 days a week without interruption. They can access it from any part of the world with proper internet.
NFR-6	**Scalability**	The application will be able to handle a rise in the no.of users & generate higher versions.

PROJECT DESIGN

Data Flow Diagram



Solution Architecture

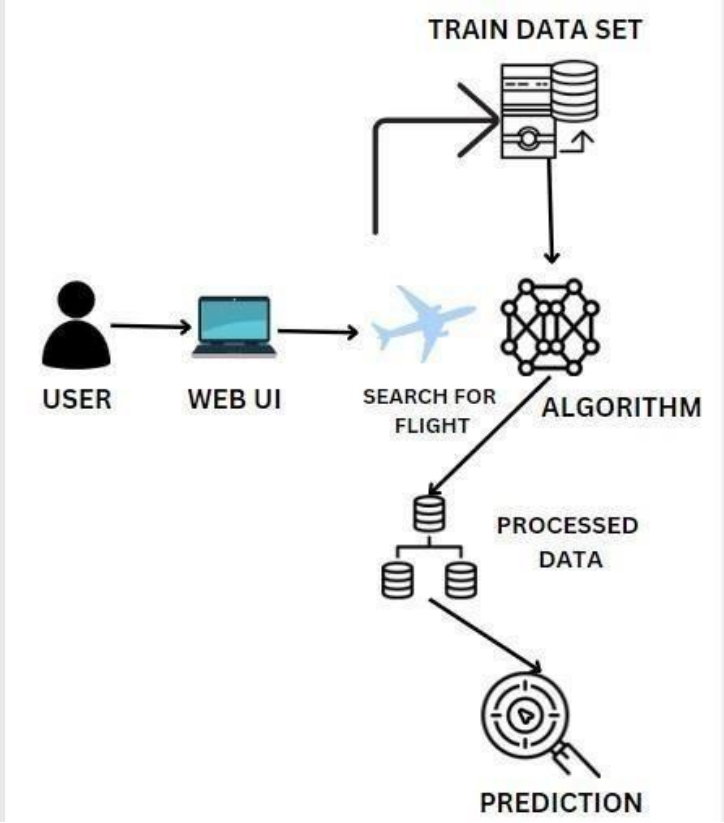
Solution architecture is a complex process – with many sub-processes – that bridges the gap between business problems and technology solutions. Its goals are to:

- Find the best tech solution to solve existing business problems.

- Describe the structure, characteristics, behavior, and other aspects of the software to project stakeholders.
- Define features, development phases, and solution requirements.

- Provide specifications according to which the solution is defined, managed, and delivered.

Solution Architecture Diagram:



User Stories

User Type	Functional requirement	User story number	User story/task	Acceptance criteria	Priority	Release
Customer (Mobile user, Web user, Care executive,	Registration	USN-1	As a user, I can register for the application by entering	I can access my account/ dashboard	High	Sprint-1

Administrator)			my mail, password, and confirming my password			
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
	Dashboard	USN-3	As a user, I can register for the application through internet	I can register & access the dashboard with Internet login	Low	Sprint-2
			application through Gmail can log into	Gmail Log in with my id	Medium	Sprint-1
			application through Gmail	Gmail		
	Login	USN-5	As a user, I can log into	I can login with my id	High	Sprint-1

			the application by entering email & password	and password		

PROJECT PLANNING &SCHEDULING

Sprint Planning & Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	2	High	Preethi S
Sprint-1	User Confirmation	USN-2	As a user, I will receive confirmation email once I have registered for the web application.	1	Medium	Madhubala A Rohini S
Sprint-1	Login	USN-3	As a user, I can login to the application by entering my email & password.	1	Medium	Akshaya S

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	4 Days	24 Oct 2022	28 Oct 2022	20	30 Oct 2022
Sprint-2	20	5 Days	28 Oct 2022	02 Nov 2022	20	05 Nov 2022
Sprint-3	20	8 Days	02 Nov 2022	10 Nov 2022	20	12 Nov 2022
Sprint-4	20	9 Days	10 Nov 2022	19 Nov 2022	20	20 Nov 2022

CODING & SOLUTIONING

Feature 1

- IBM Watson Platform
- Node red
- Web UI

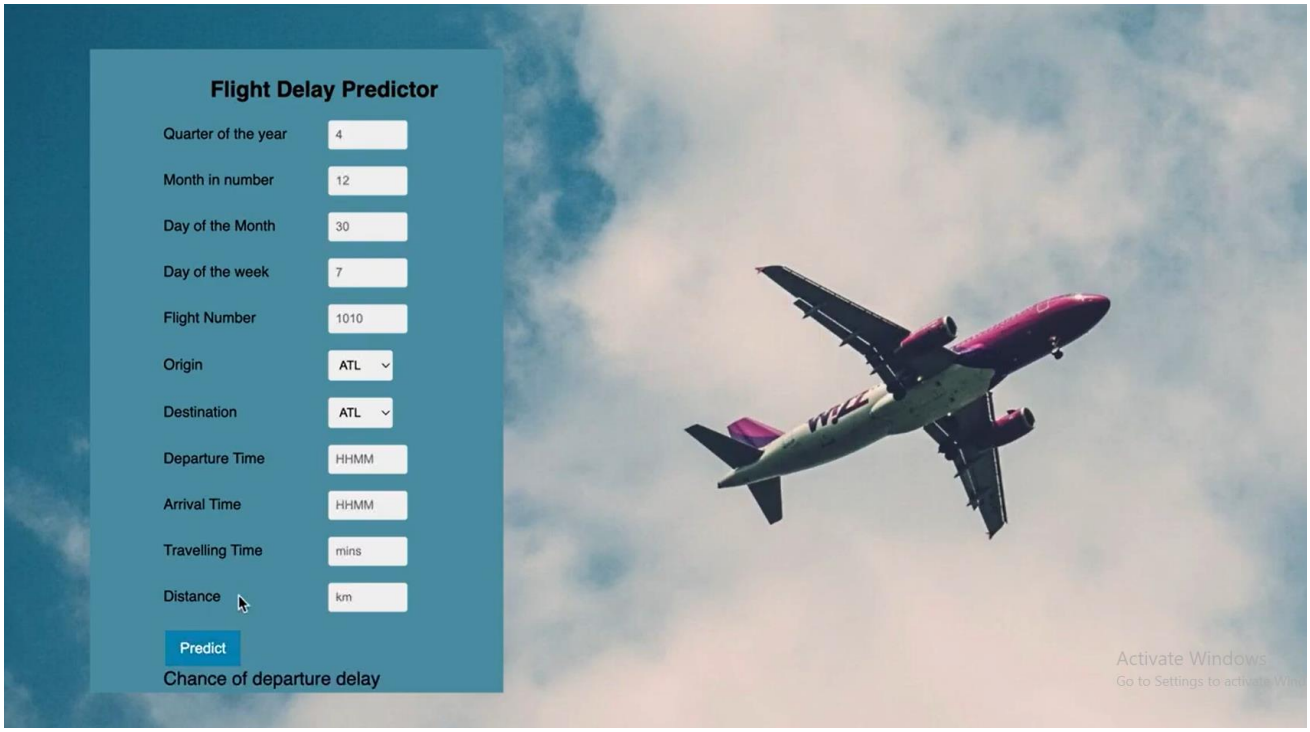
- Python code-Flask
- HTML
- CSS

Feature 2

- Data entry page
- Prediction result page

TESTING AND RESULTS

Test Cases Test case 1:



ADVANTAGES

- Machine learning can predict flight delays with a high degree of accuracy.
- Machine learning can help identify causes of flight delays.
- Machine learning can help reduce the number of flight delays.
- Machine learning can help improve the efficiency of airport operations.

DISADVANTAGES

- Machine learning models can be complex and difficult to understand.
- Machine learning models require a large amount of data to train and can be time-consuming to develop.
- Machine learning models can be prone to overfitting, meaning they may not generalize well to new data.
- Machine learning models can be expensive to develop and maintain.

CONCLUSION

In this project, we use flight data, weather, and demand data to predict flight departure delay. Our result shows that the Logistic Regession yields the best performance compared to the SVM model. Somehow the SVM model is very time consuming and does not necessarily produce better results. In the end, our model correctly predicts 91% of the non-delayed flights. However, the delayed flights are only correctly predicted 41% of time. As a result, there can be additional features related to the causes of flight delays that have not yet been discovered using our existing data sources.

In the second part of the project, we can see that it is possible to predict flight delay patterns from just the volume of concurrently published tweets, and their sentiment and objectivity. This is not unreasonable; people tend to post about airport delays on Twitter; it stands to reason that these posts would become more frequent, and more profoundly emotional, as the delays get worse. Without more data, we cannot make a robust model and find out the role of related factors and chance on these results. However, as proof of concept, there is potential for these results. It may be possible to routinely use tweets to ascertain an understanding of concurrent airline delays and traffic patterns, which could be useful in a variety of circumstances.

FUTURE SCOPE

This project is based on data analysis from the year 2008. A large dataset is available from 1987-2008 but handling a bigger dataset requires a great amount of preprocessing and cleaning of the data. Therefore, the future work of this project includes incorporating a larger dataset. There are many ways to preprocess a larger dataset like running a Spark cluster over a server or using

cloud-based services like AWS and Azure to process the data. With the new advancement in the field of deep learning, we can use Neural Networks algorithm on the flight and weather data. Neural Network works on the pattern matching methodology. It is divided into three basic parts for data modelling that includes feed forward networks, feedback networks, and self-organization network. Feed-forward and feedback networks are generally used in the areas of prediction, pattern recognition, associative memory, and optimization calculation, whereas self-organization networks are generally used in cluster analysis. Neural Network offers distributed computer architecture with important learning abilities to represent nonlinear relationships.

Also, the scope of this project is very much confined to flight and weather data of United States, but we can include more countries like China, India, and Russia. Expanding the scope of this project, we can also add the flight data from international flights and not just restrict ourselves to domestic flights.

Github Link:

https://github.com/IBM-EPBL/IBM-Project-1225-1658379233.git

PROJECT DEMO LINK:

https://youtu.be/KOxDaN-BDys