Project Report

Developing a Flight delay prediction model using machine learning

Team members

1.Ahamadullah Javith

2.Hariharan

3.Ameen

4. Asraf ali

Project Report Format

Introduction

1.1 Project Overview

As people increasingly choose to travel by air, the amount of flights that fail to take off on time also increases. This growth exacerbates the crowded situation at airports and causes financial difficulties within the airline industry. Air transportation delay indicates the lack of efficiency of the aviation system. It is a high cost to both airline companies and their passengers. Therefore, predicting flight delays can improve airline operations and passenger satisfaction, which will result in a positive impact on the economy. Our main goal is to accurately predict the flight delays in order to increase customer satisfaction and income of airlines.

1.2 Purpose

The main purpose of our flight delay prediction model is to accurately predict flight delays with certain features including airlines who operate them, distance they have to cover, origin airport, target airport, departure times and so on. Being able to accurately predict flight delays can help the passengers know what delay should be ready to face depending on where they fly from and the airlines they chose to fly.

Literature survey

2.1 Existing solution

Nowadays, service quality plays an important role in attracting customers. Among these, air travels have their special customers and the most important matter in these travels is the flight time, on time arrival at destination for passengers such those who have an important meeting, that has been leading to high expenses for the passenger until get to their destination on time. Flight delay has negative economic effects on the passenger, agencies and airports. Therefore, any reduction of these effect requires decreasing postponed flight price, so that prediction or estimation has a great significance and numerous studies has been to dedicated this subject. Correspondingly, all the scientists have tried to design a model that understands effective factors and computes effect of each factor and their relation. Overall, the prediction methods are classified into five groups including Statistical Methods, Probability methods, network-based methods, operational methods and machine learning methods. In one of the best studies that has been performed based on statistics delay time has been considered to be reduced. Their study has investigated important factors before fly and those which occur on the ground. In the next step, it has predicted the delay at destination based on factors that occur in the vicinity of arrival time at destination. Eventually, results have shown that whenever, the delay is correctly predicted, pas-senger disaffection and fuel consumption decrease and consequently number of flight increases. Moreover, it is possible to increase the agencies' benefits through reducing number of passengers who wrongly selected their routs or specifying the probabilities for some flights and optimizing delay time prediction. Another prominent investigation based on Probability has been done and the author believes that huge storm in U.S.A has highly affected the flight delay. This study has been devoted to predict delay based on mathematical calculations and through con-sidering delay time duration of the flights that had been engaged to storm in the same day. Metrological reports have shown the effect of storm one hour before and after event cause ephemeral climate at the region. In the next step, Monte Carlo simulation has been used to estimate the airport runway capacity, so that traffic of each runway would have been estimated. As the research has employed only one factor, the model has not enough accuracy, but it is possible to increase region air capacity path structure. A model has been presented in , which is one of the best network-based mod- els. The researchers have presented a model based on Bayesian and Gaussian mixture model- expectation maximization (GMM-EM) algorithm to predict and analyze the factors affecting the flight delay in Brazil for several point along the path. At the first stage of model, the degree of effectiveness for each factor is specified and then it has speci- fied investigated whether the delay had happened in a greater domain or no. the next delay probability is computed using GMM-EM and EM algorithm which are speci- fied based on similarity. The result has shown that it is possible to predict the probability of delay in higher levels through specifying low level factors. Moreover, GMM EM similarity function has more values rather than EM algorithm in each step, so that the results would have been converged sooner. In addition, the model accuracy is increased, so that the prediction is more trustable. One of the best studies in the area of operating method has been presented. Studied the effects of capacity and damage on different levels of delay in American airports. Other simulations focus on stability and reliability during the delay and its propaga- tion. For instance, inthe problems of congestion were studied. Then, a queue-based model was presented for analyzing delay propagation in consecutive flights in the Los Angeles airport.

2.2 References

1. Kuhn, Nathalie and Navaneeth Jamadagni. "Application of Machine Learning Algorithms to Predict Flight Arrival Delays." (2017).

2. N, Prabakaran & Kannadasan, Rajendran. (2018). Airline Delay Predictions using Supervised Machine Learning. International Journal of Pure and Applied Mathematics. 119.

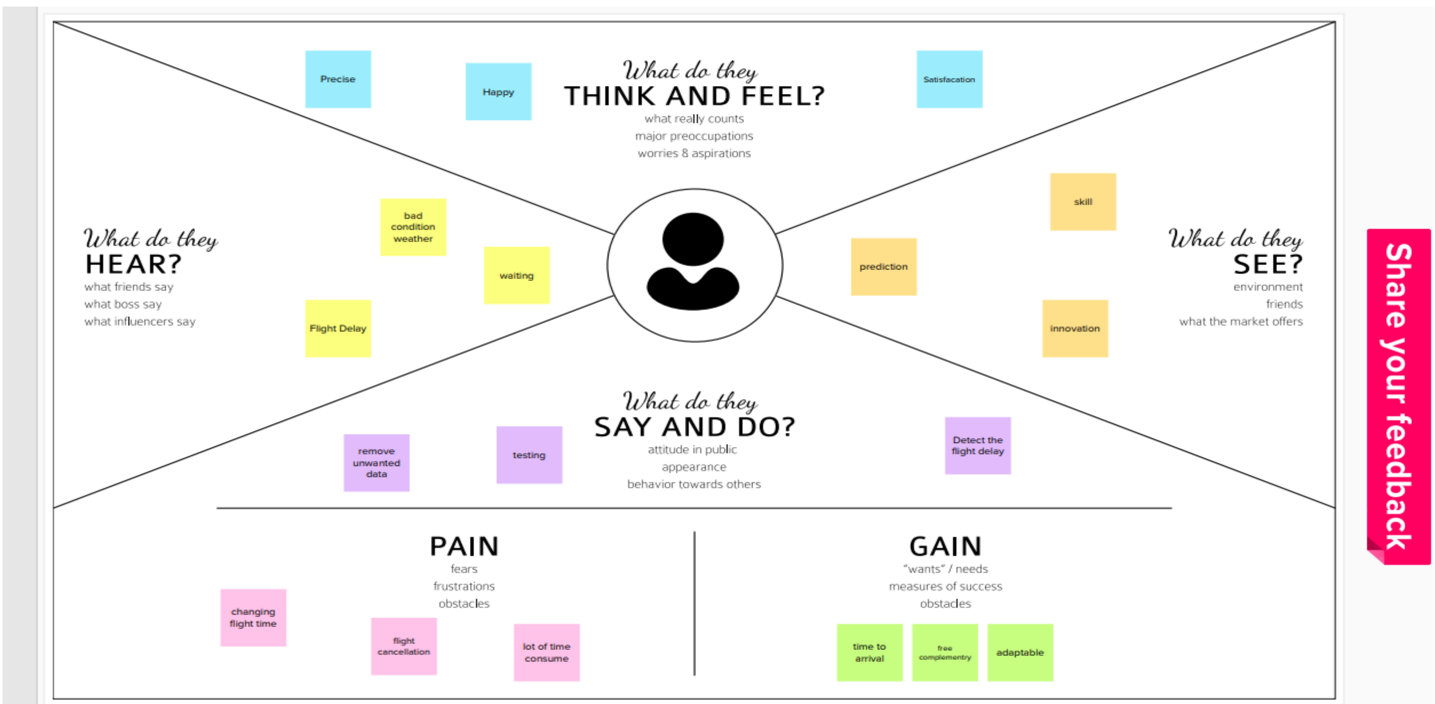
3. Yuemin Tang. 2021. Airline Flight Delay Prediction Using Machine Learning Models. In *2021 5th International Conference on E-Business and Internet (ICEBI 2021), October 15-17, 2021, Singapore, Singapore*. ACM, New York, NY, USA, 7 Pages. **https://doi.org/10.1145/3497701.3497725**

2.3 Problem statement definition

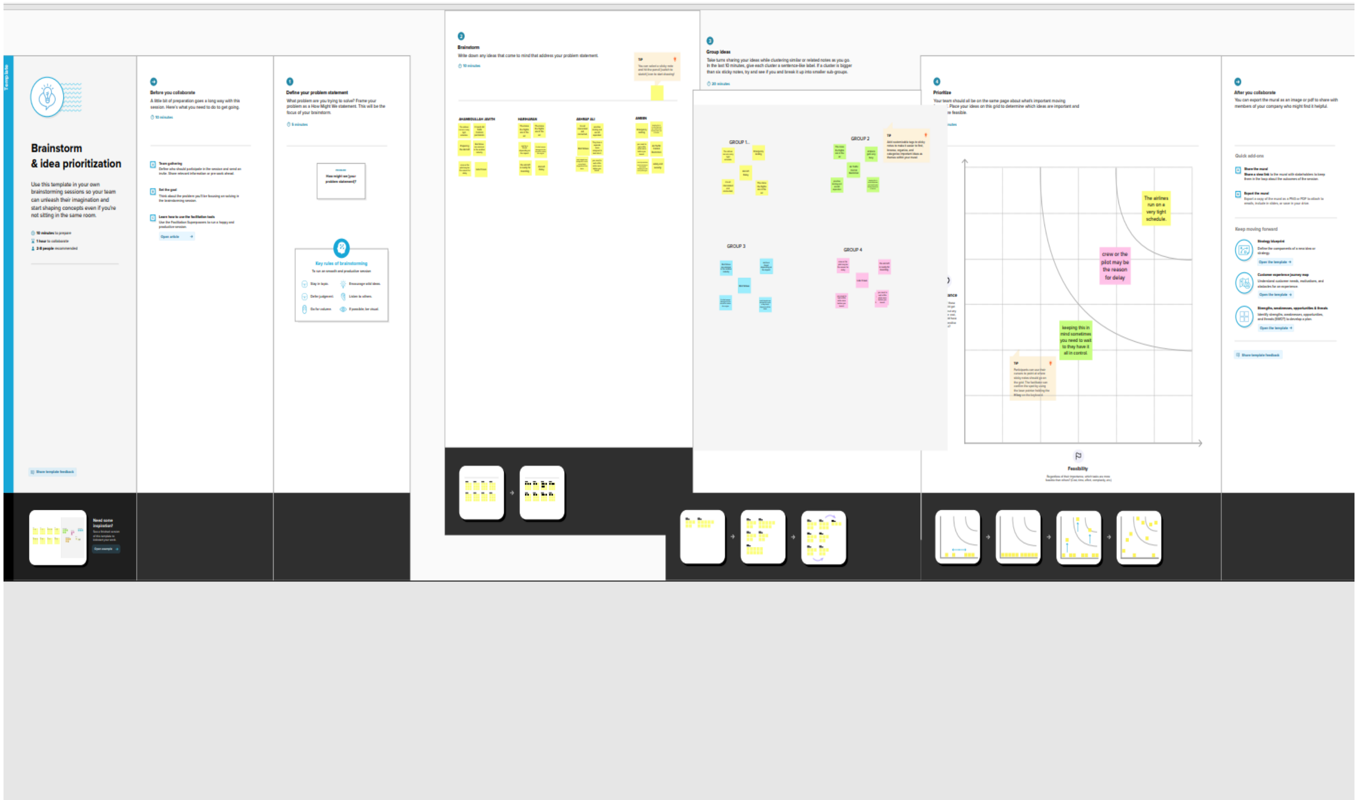
Flight delays represent a common problem in everyday air traffic practice. The impact of flight delay can be a risk and this risk represents financial losses, the dissatisfaction of passengers, time losses, loss of reputation and bad business relations. Flight delays not only irritate air passengers and disrupt their schedules but also cause a decrease in efficiency, an increase in capital costs, reallocation of flight crews and aircraft, and additional crew expenses. o overcome the gradually increasing flight delay we're using supervised machine learning algorithms in order to predict accurate delays in flights which allow passengers to be well prepared for their journey and enables airlines to respond to various cause of flight delay in advance.

Idea and proposed solution

3.1 Empathy Map canvas



3.2 Ideation and brainstorming



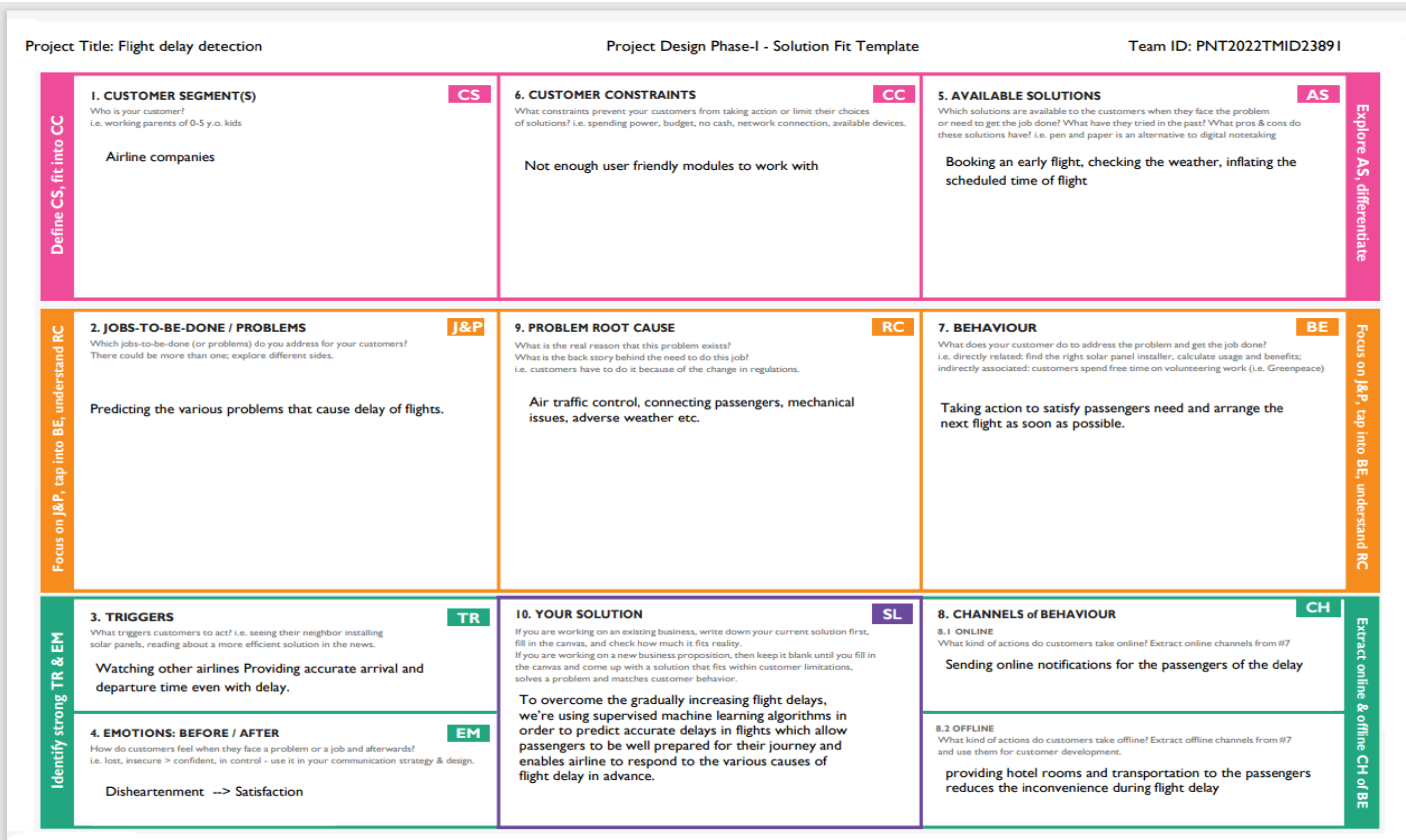
3.3 Proposed solution

Proposed Solution Template:

Project team shall fill the following information in proposed solution template.

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	Issues including late arrival of aircraft, mechanical delays, waiting for connecting passengers or bags or extreme weather events can cause flight delays or led to flight cancellations in some cases.
2.	Idea / Solution description	To overcome the gradually increasing flight delay we're using supervised machine learning algorithms in order to predict accurate delays in flights which allow passengers to be well prepared for their journey and enables airlines to respond to various cause of flight delay in advance.
3.	Novelty / Uniqueness	Provides appropriate departure and arrival time which helps in passengers easy travelling.
4.	Social Impact / Customer Satisfaction	Flight delay not only irritates air passengers and disrupts their schedules but also cause a decrease in efficiency, increase in capital and additional crew expenses.
5.	Business Model (Revenue Model)	Monthly subscription
6.	Scalability of the Solution	To make the system more scalable supervised machine learning algorithm is used in order to provide better accuracy.

3.4 Problem Solution fit


Project Title: Flight delay detection
Project Design Phase-I - Solution Fit Template
Team ID: PNT2022TMID23891

Define CS, fit into CC

1. CUSTOMER SEGMENT(S) CS
Who is your customer?
i.e. working parents of 0-5 y.o. kids
Airline companies

6. CUSTOMER CONSTRAINTS CC
What constraints prevent your customers from taking action or limit their choices of solutions? i.e. spending power, budget, no cash, network connection, available devices.
Not enough user friendly modules to work with

5. AVAILABLE SOLUTIONS AS
Which solutions are available to the customers when they face the problem or need to get the job done? What have they tried in the past? What pros & cons do these solutions have? i.e. pen and paper is an alternative to digital notetaking
Booking an early flight, checking the weather, inflating the scheduled time of flight

Explore AS, differentiate

Focus on J&P, tap into BE, understand RC

2. JOBS-TO-BE-DONE / PROBLEMS J&P
Which jobs-to-be-done (or problems) do you address for your customers? There could be more than one; explore different sides.
Predicting the various problems that cause delay of flights.

9. PROBLEM ROOT CAUSE RC
What is the real reason that this problem exists?
What is the back story behind the need to do this job?
i.e. customers have to do it because of the change in regulations.
Air traffic control, connecting passengers, mechanical issues, adverse weather etc.

7. BEHAVIOUR BE
What does your customer do to address the problem and get the job done?
i.e. directly related: find the right solar panel installer, calculate usage and benefits; indirectly associated: customers spend free time on volunteering work (i.e. Greenpeace)
Taking action to satisfy passengers need and arrange the next flight as soon as possible.

Focus on J&P, tap into BE, understand RC

Identify strong TR & EM

3. TRIGGERS TR
What triggers customers to act? i.e. seeing their neighbor installing solar panels, reading about a more efficient solution in the news.
Watching other airlines Providing accurate arrival and departure time even with delay.

4. EMOTIONS: BEFORE / AFTER EM
How do customers feel when they face a problem or a job and afterwards?
i.e. lost, insecure > confident, in control - use it in your communication strategy & design.
Disheartenment --> Satisfaction

10. YOUR SOLUTION SL
If you are working on an existing business, write down your current solution first, fill in the canvas, and check how much it fits reality.
If you are working on a new business proposition, then keep it blank until you fill in the canvas and come up with a solution that fits within customer limitations, solves a problem and matches customer behavior.
To overcome the gradually increasing flight delays, we're using supervised machine learning algorithms in order to predict accurate delays in flights which allow passengers to be well prepared for their journey and enables airline to respond to the various causes of flight delay in advance.

8. CHANNELS of BEHAVIOUR CH
8.1 ONLINE
What kind of actions do customers take online? Extract online channels from #7
Sending online notifications for the passengers of the delay

8.2 OFFLINE
What kind of actions do customers take offline? Extract offline channels from #7 and use them for customer development.
providing hotel rooms and transportation to the passengers reduces the inconvenience during flight delay

Extract online & offline CH of BE

4.REQUIREMENT ANALYSIS

4.1 Functional Requirement

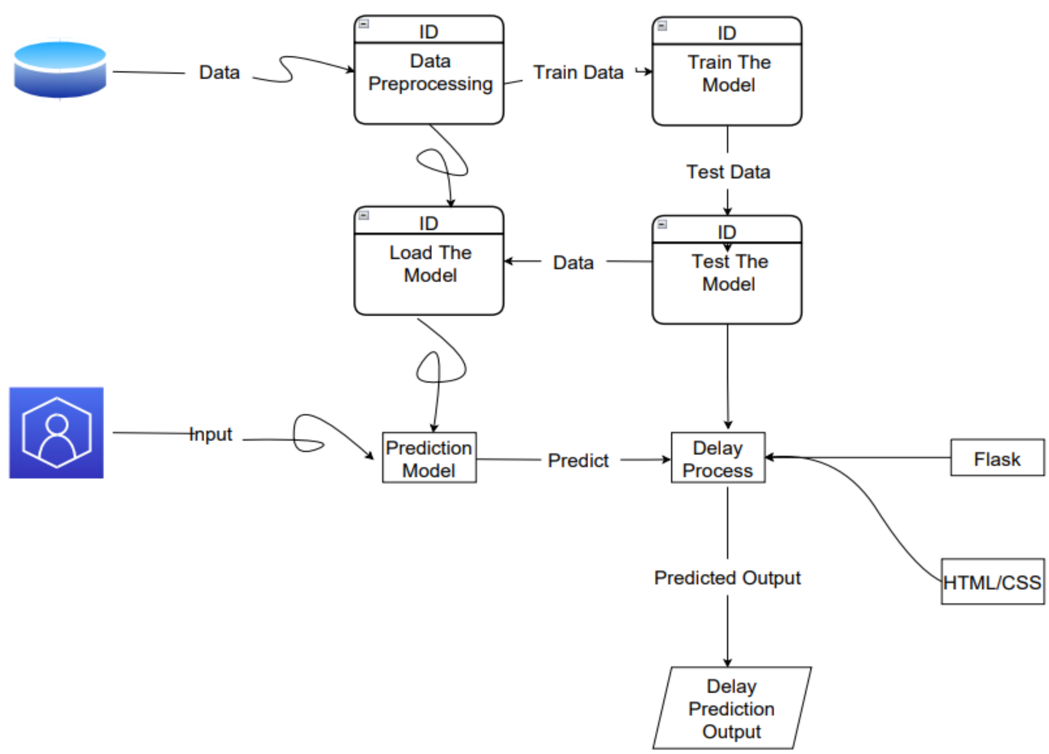
FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration through Gmail
FR-2	User Confirmation	Confirmation via Email
FR-3	User login	Login with Gmail and Password
FR-4	Data input	Enter the necessary input
FR-5	Data output	View the output for the entered input
FR-6	Compliance to law	To meet regulatory compliance and standard requirement

4.2 Nonfunctional requirement

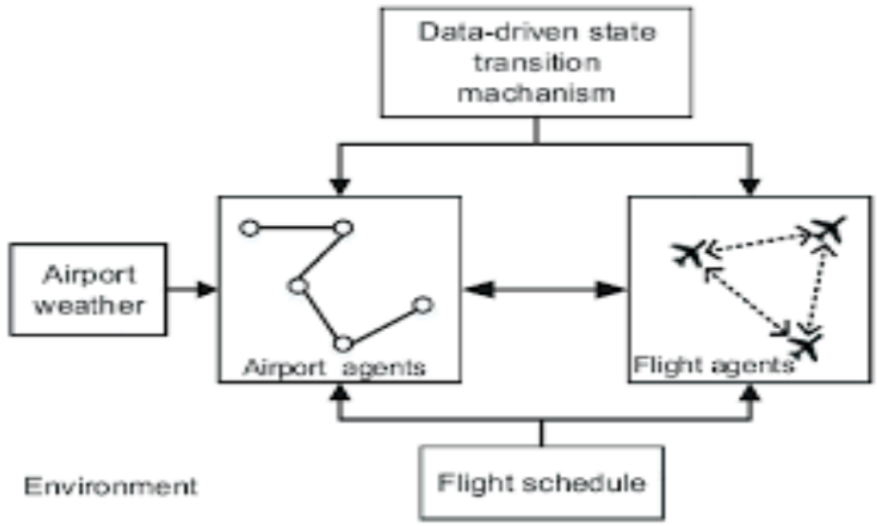
FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	The UI of the program is diagram oriented which is easy to use
NFR-2	**Security**	Only administrator can give access to the desired systems
NFR-3	**Reliability**	Previous model is saved as a backup in case the recent model fails.
NFR-4	**Performance**	The UI of the program must load faster and the result must be given faster
NFR-5	**Availability**	New module must not impact the UI of the program. Availability mustn't take longer
NFR-6	**Scalability**	It should be scalable enough to support multiple users at a time to avoid web traffic.

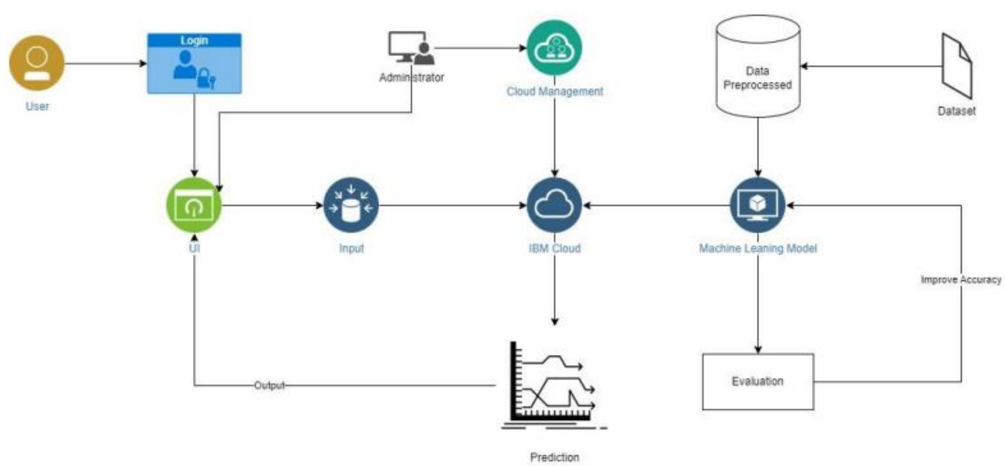
Project Design

5.1 Data flow diagrams



5.2 Solution and technical architecture





5.3 User stories

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority
Customer (Mobile user)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account / dashboard	High
		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
		USN-3	As a user, I can register for the application through Facebook	I can register & access the dashboard with Facebook Login	Low
	Login	USN-4	As a user, I can register for the application through Gmail	I can access my dashboard	Medium
	Dashboard	USN-5	As a user, I can log into the application by entering email & password	I can plan according to flight delay	High
Customer (Web user)	Technical Support	USN-6	As a user, I have to rectify any errors that occur in the program.	I can access the program files.	High
Customer Care Executive	Data Update	USN-7	As a user, I have to update the dataset with the recent flight details	I can collect and verify the data	Medium
Administrator	Maintenance	USN-8	As a user, I have to maintain the program for any updates	I can maintain the program	Low

Project planning and scheduling

6.1 Sprint Planning & Estimation

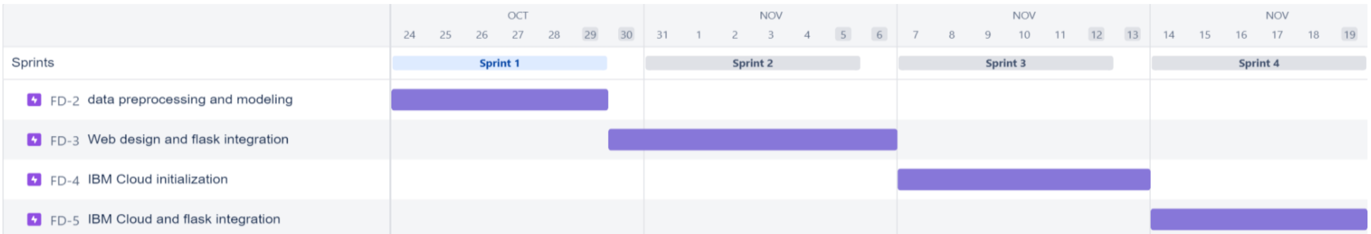
Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Data Collection and Pre-processing	USN-1	As a user, I can't interact anything. Waiting is user's task. User can listen the relationship exist between the various attributes of data by presentation of developer	2	High	Ahamedullah Javith, Ameen
Sprint-1	Model Building	USN-2	As a user, I can predict flight delay by various developed ML models by console	1	High	Ahamedullah Javith, Ameen

Sprint-2	Model Evaluation	USN-3	As a user, I can predict flight delay by best Model in various developed ML model by console	2	High	Ahamedullah Javith, Ameen
Sprint-2	Model Deployment on IBM Cloud using IBM Watson	USN-4	As a user, I can use the model by requesting the deployed model on Cloud	1	Medium	Ahamedullah Javith Ameen
Sprint-2	Basic user interaction Dashboard	USN-5	As a user, I can use the model or prediction from model by interacting with dashboard	2	High	Hariharan, Ashraf Ali
Sprint-3	Improved Dashboard and GUI	USN-6	As a user, I can use the model or prediction from model by interacting with improved dashboard	1	Medium	Hariharan, Ashraf Ali
Sprint-3	Registration	USN-7	As a user, I can register for the application by entering my email, password, and confirming my password.	2	High	Ashraf Ali, Ahamedullah Javith
Sprint-3	Registration	USN-7	As a user, I can register for the application by entering my email, password, and confirming my password.	2	High	Hariharan, Ashraf Ali
Sprint-3	Login	USN-8	As a user, I can log into the application by entering email & password and I can register .login to the application through Gmail	2	Medium	Ameen, Hariharan
Sprint-4	Raise query/complaint and give feedback	USN-9	As a user, I can raise complaint or query and give feedback	1	Medium	Ashraf Ali, Hariharan
Sprint-4	Improve overall web app	USN-10	As a user, I can user revised and improved version of web application	1	High	Ahamedullah Javith, Ameen, Ashraf Ali, Hariharan

6.2 Sprint delivery schedule

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	6 Days	24 Oct 2022	29 Oct 2022	20	31 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	20	07 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	20	12 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	20	19 Nov 2022

6.3 Reports for JIRA



7. CODING & SOLUTIONING

We started by preprocessing the data and then visualized the data to see patterns. Then we used Decision Tree as our model as it provided better prediction accuracy. We further tuned the model by hyper parameter tuning for our model using Grid Search CV. In Grid Search CV we can see that max depth of 6 and min samples split of 2 provided the best accuracy and then we cross validated the model using KFold function with the k value as 6 which gave an accuracy of 92%.

```
In [31]: from sklearn.model_selection import cross_val_score, KFold, GridSearchCV
         KF = KFold(n_splits =6, shuffle = True, random_state = 25)
         params={'max_depth':[4,5,6],
                'min_samples_split':[2,3,4],
                'criterion':['gini','entropy','log_loss']}

In [35]: grid_cv=GridSearchCV(classifier,params,cv = KF)
         grid_cv.fit(X_train,y_train)

         ...

In [36]: grid_cv.best_params_

Out[36]: {'criterion': 'entropy', 'max_depth': 4, 'min_samples_split': 2}

In [38]: cv_results=cross_val_score(classifier,X_train,y_train,cv=KF)
         print(cv_results)

         [0.89604344 0.92242048 0.92546584 0.90993789 0.92701863 0.92934783]

In [ ]:
```

8. TESTING

1.Test Cases

				Maximum Marks	4 marks					
Test case ID	**Feature Type**	**Component**	**Test Scenario**	**Pre-Requisite**	**Steps To Execute**	**Test Data**	**Expected Result**	**Actual Result**	**Status**	**Commnets**
prediction_TC_001	Functional	Home Page	Verify user is able to see the prediction input page when user clicked on url	chrome browser	1.Enter URL and click go 2.Verify prediction displayed or not	flask app	prediction input page should display	Working as expected	Pass	
prediction_TC_002	UI	Home Page	Verify the user is able to predick the flight with the proper details	chrome browser	1.Enter URL and click go 2..Verify prediction with below UI elements: a.flight number b.date c.orgin and destination dropdownbox d.flights timing e.prediction button	flask app	Application should be shown below a.flight number b.date c.orgin and destination dropdownbox d.flights timing e.prediction button	Working as expected	Pass	
prediction_TC_003	Functional	Home page	Verify user is able to log into application with InValid input	chrome browser	1.Enter URL and click go 2.Enter the flight number 3.Enter the date 4.Enter the orgin and destination dropdownbox 5. Enter the flights timing 6.Enter the prediction button	flight number:23587 month:12 day :12 orgin : ALT Destinsation:SEA sheduled dept time:1215 Actual dept time:1236	User should navigate to result page recived properly		Pass	
prediction_TC_004	Functional	Login page	Verify user is able to log into application with InValid input	chrome browser	1.Enter URL and click go 2.Enter the flight number 3.Enter the date 4.Enter the orgin and destination dropdownbox 5. Enter the flights timing 6.Enter the prediction button	flight number:23587 month:12 day :12 orgin : ALT Destinsation:Alt sheduled dept time:1215 Actual dept time:1236 sheduled Arr Time:1420	Application should show 'the orgin and destination cannot be samevalidation message.	Working as expected	Fail	the orgin and destination cannot be same
prediction_TC_005	Functional	Login page	Verify user is able to log into application with InValid credentials	chrome browser	1.Enter URL and click go 2.Enter the flight number 3.Enter the date 4.Enter the orgin and destination dropdownbox 5. Enter the flights timing 6.Enter the prediction button	flight number:23587 month:13 day :12 orgin : ALT Destinsation:SEA sheduled dept time:1215 Actual dept time:1236 sheduled Arr Time:1420	Application should show the month value more than 12 validation message.	Working as expected	Fail	the month value more than 12
Result_TC_001	UI	result page	Verify user is able toview predic	chrome browser	1.Enter URL and click go 2.Enter the orrect input values and click the prediction button	flight number:23587 month:12 day :12 orgin : ALT Destinsation:SEA sheduled dept time:1215 Actual dept time:1236 sheduled Arr Time:1420	Application should show flight is	as expected	Pass	

2. User Acceptance Testing

2. Defect Analysis

This report shows the number of resolved or closed bugs at each severity level, and how they were resolved

Resolution	Severity 1	Severity 2	Severity 3	Severity 4	Subtotal
By Design	7	5	2	1	15
Duplicate	0	0	0	0	0
External	5	0	0	1	6
Fixed	11	2	4	20	37
Not Reproduced	0	0	1	0	1
Skipped	0	0	0	0	0
Won't Fix	0	0	0	0	0
Totals	23	7	7	23	59

3. Test Case Analysis

This report shows the number of test cases that have passed, failed, and untested

Section	Total Cases	Not Tested	Fail	Pass
Model evaluation	10	0	0	10
Client Application	20	0	0	20
Exception reporting	5	0	0	5
Final Report Output	2	0	0	2

9. Results

9.1 Performance metrics

```
In [27]: from sklearn.metrics import accuracy_score,confusion_matrix, classification_report
         print(accuracy_score(y_test, pred))

         0.923332327195895

In [28]: print(confusion_matrix(y_test, pred))

         [[2809  126]
          [ 128  250]]

In [29]: print(classification_report(y_test, pred))

                       precision    recall  f1-score   support

                  0.0       0.96      0.96      0.96      2935
                  1.0       0.66      0.66      0.66       378

             accuracy                           0.92      3313
            macro avg       0.81      0.81      0.81      3313
         weighted avg       0.92      0.92      0.92      3313

In [ ]:
```

10. Advantages and Disadvantages

Advantages: The flight delay prediction model is user friendly and easy to interact .our model predicts overall 92% accuracy so that it can be used to provide better customer experience and reduce overall expenses.

Disadvantages: This model is limited to only fix number of Airports on those particular surroundings so; flight delay on other airports can't be predicted.

11. Future works

This project is limited to only a certain place for now but in future we can expand this project by adding more countries so that the flight delay can be minimized all around the world. Thus expanding the scope of this project, we can also add the flight data from international flight and not just restrict our self to domestic flights.

12. Conclusion

Predicting flight delay is on interesting research topic and required many attentions these years. Majority of research have tried to develop and expand their models in order to increase the precision and accuracy of predicting flight delays. In our project, by using machine learning algorithm our flight delay prediction model can predict 92% accuracy. When comparing decision tree algorithm with other algorithms like KNN, SVM, Logistic regression etc. Decision tree algorithm gives better performance and obtained the highest score compared to all other algorithms. For Flight data we have used Scheduled departure time, scheduled arrival time, actual destination time, destination, origin By applying these data's one could be able to predict whether a flight might be delayed, and more importantly, how long delayed time she/he would expect. However, there is some limitation in our model, first, our model only included limited capability, as more years of data included, the prediction could be easier. In addition, some other related information such as airplane type, e.g., detailed weather data specific to airport was not included.

Therefore, we will try to collect more related data and deploy better computational powers to build a better model in future.

13. APPENDIX

Source code

```
import sys
import numpy as np
import pandas as pd
import seaborn as sns
import pickle
%matplotlib inline
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import OneHotEncoder
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
import sklearn.metrics as metrics
dataset =pd.read_csv("C:\\Users\\javith\\Downloads\\flightdata.csv")
dataset.head()
dataset.drop('Unnamed: 25', axis = 1, inplace = True)
dataset.info()
df.isnull().sum()
dataset.dropna(subset=['DEP_TIME','ARR_DELAY'], inplace = True)
dataset.isnull().sum()
dataset.shape
ax = sns.countplot(y = dataset['ORIGIN'], order =
dataset['ORIGIN'].value_counts().index);
ax.set_title("Airports w.r.t Depature Flights", fontsize = 16);
ax.set_xlabel("Number of Flights", fontsize = 14);
ax.set_ylabel("Airport Code", fontsize = 14);
ax.bar_label(ax.containers[0], label_type = 'center', color = 'white', size = 14);
ax = sns.countplot(y = dataset['DEST'], order = dataset['DEST'].value_counts().index);
ax.set_title("Airports w.r.t Arrival Flights", fontsize = 16);
ax.set_xlabel("Number of Flights", fontsize = 14);
ax.set_ylabel("Airport Code", fontsize = 14);
ax.bar_label(ax.containers[0], label_type = 'center', color = 'white', size = 14);
fig, ax = plt.subplots(1, 2, figsize = (10,10))
```

```
ax[0].pie(dataset['DEP_DEL15'].value_counts(), labels = ['On Time', 'Delayed'],
autopct = '%1.2f%%',startangle = 90, explode = (0,0.1));
ax[0].title.set_text("Ratio of Delayed Departure Flights");
ax[1].pie(dataset['ARR_DEL15'].value_counts(), labels = ['On Time','Delayed'],
autopct = '%1.2f%%',startangle = 90, explode = (0,0.1));
ax[1].title.set_text("Ratio of Delayed Arrival Flights");
sns.heatmap(dataset.corr());
new_dataset = pd.get_dummies(dataset, columns = ['ORIGIN','DEST'])
new_dataset.head()
X =
new_dataset[['MONTH','DAY_OF_MONTH','DAY_OF_WEEK','ORIGIN_ATL','ORIGIN_DT
W','ORIGIN_JFK','ORIGIN_MSP','ORIGIN_SEA','DEST_ATL','DEST_DTW','DEST_JFK','DES
T_MSP','DEST_SEA','CRS_DEP_TIME','DEP_TIME','DEP_DEL15','CRS_ARR_TIME']]
y = new_dataset['ARR_DEL15']
X.head()
y.head()
X.shape
y.shape
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30)
from sklearn.tree import DecisionTreeClassifier
Classifier = DecisionTreeClassifier(max_depth = 4, min_samples_split = 4,
random_state = 25)
Classifier.fit(X_train, y_train)
from sklearn.metrics import accuracy_score
print(accuracy_score(y_test, pred))
Classifier.predict([[1,4,1,0,1,0,0,0,0,0,0,0,1,1215,1236,1,1420]])
import pickle
pickle.dump(Classifier, open('flightClassifier.pkl','wb'))
```

github link: https://github.com/IBM-EPBL/IBM-Project-15598-1659601321

demo

link:https://drive.google.com/file/d/1qfs6zMuJbJGQTTim6XjNwKJRPZJus3bW/view?usp=drivesdk