ABSTRACT

This project's main objective is to forecast aeroplane delays brought on by many factors. Negative effects of flight delays are mostly financial for travellers, the airline industry, and airport authorities. Additionally, in the area of sustainability, it even has the potential to harm the environment due to an increase in fuel consumption and gas emissions. Therefore, these considerations show how important and crucial it has become to estimate delays, regardless of the several airline meshes. To perform the predictive analysis, which entails a variety of statistical methods from supervised machine learning and data mining, that examines recent and historical data to make predictions or simply analyse about upcoming delays, with the aid of Python 3's Regression Analysis and regularisation technique.Here, a forecasting method for aeroplane delays is based on weather-related potential delays. The primary variables used in the technique for predicting delays are the temperature, humidity, millimetres of rain, visibility, and month.

CHAPTER 1
INTRODUCTION

The aviation networks expand daily as the number of flights increases.complex. This enables us to reach more places and do tasks in various places.shorter intervals of time. Managing all of these operations can occasionally prove challenging and This increases the possibility of a flight delay. Flight delays may have detrimental consequences on airline customers, staff members, the economy, and the environment. The goal of this research is to develop a model to forecast aeroplane departure delays. There are many variables that can affect a flight delay, including those relating to an airplane's characteristics, the airport's characteristics, the number of passengers and cargo, the preparation processes chosen for the aircraft, the weather, the aircraft's prior flights, the airline's and the airport's operational intensity. Additional variables can be added to this list. As a result, the party conducting the study may choose from a variety of features as an input to build a model relating to priorities and the degree of control over elements.

This study examines domestic commercial flights in the United States in August 2018. Additionally, flight data is coupled with information about aircraft, passenger boarding, and cargo to potentially provide insights into these issues.The Bureau of Transportation Statistics' website pages were used to collect all the data. Supervised machine learning techniques are employed to forecast flight delays, and the results are examined. These techniques include decision trees, random forests, bagging classifiers, additional trees classifiers, gradient boosting, and xgboost classifiers.

CHAPTER 2
LITERATURE SURVEY

A.LOGISTIC REGRESSION MODEL

No data mining projects could be finished without thoroughly understand the data first. So, in order to better understand data, we start our project by exploring the data first. We found the original dataset includes 28 attributes/columns and while most of the data were in float format, some of them were object types). In addition, as shown, there were also many null values in the original datasets. So, we need to first clean the columns with null values and change data types of objects into suitable types (mostly integers) for the convenience of machine learning.

B.LINEAR REGRESSION MODEL

We renamed the original data column names and validated the nulls, however with a little different approach. We first plotted a density plot for chosen attributes. After plotting the density plot with columns with “nan” values, we found none of the columns strictly follows normal distribution, and most of them were largely skewed and concentrated to only few values. Replacing methods, we tried included applying fill na() method to replace “nan” and replacing missing “nan” values with the mean of corresponding columns. However, none of the methods enable us to develop model with desirable results. So instead of replacing “nan” with normal distribution, we decided to use merely replace “nan” with extreme values that without the original data range.

C.INITIAL DATA EXPLORING

After data cleaning we start the first process of exploring our data if there were any patterns within the independent variables. The above graph shows the no of delays airline wise. On the left side you can see there is a false value, which means instances when an airline

has not been delayed. On the right-side true values suggests that there is a delay. We can see and conclude that maximum delay is caused by Southwest Airlines. Also, in the next graph we can see that the maximum number of flights are from Southwest Airlines, which compel us to think that one of the reasons for the delay is the operational process of airline. And these delays are known as career delay, we can reduce this delay with effective planning strategies.

D.DIMENSIONALITY REDUCTION

So as there were 28 columns and we wanted our model to be very precise, before going ahead we wanted to be sure there should not be any kind of correlation between the predictor variables, otherwise our model will be over fitting. So, we used correlation matrix and the criteria was, if 2 variables have correlation greater than 0.4 or less than -0.4, we will drop one of those variables. The most common correlated variables are actual departure time, actual arrival time, planned arrival time planned departure time among others, this makes sense because these factors are directly impacting the delay so there is no point of adding those variables. For example, the website FlightCaster exploit several sources of information (airports, airlines, weather and possibly historical data) to provide probabilities of being on-time, less than one hour late or more than one hour late, to travelers. However, this website is using the same estimations for all the flights when no short term information is available.

CHAPTER 3
DESCRIPTION

3.1 PROBLEM STATEMENT

Using a machine learning model, we can predict flight arrival delays. The input to our algorithm is rows of feature vector like departure date, departure delay, distance between the two airports, scheduled arrival time etc. We then use decision tree classifier to predict if the flight arrival will be delayed or not. A flight is considered to be delayed when difference between scheduled and actual arrival times is greater than 15 minutes. Furthermore, we compare decision tree classifier with logistic regression and a simple neural network for various figures of merit.

3.2 TECHNOLOGY STACK

S.No	Component	Description	Technology
1.	User Interface	User can interacts with application through Web UI.	HTML , CSS , JavaScript , Bootstrap , Flask
2.	Application Logic-1	The user can enter the data in it is sent for the machine learning model for the prediction	Java / Python
3.	Application Logic-2	The application is directly deployed in the IBM cloud	IBM Watson STT service
4.	Database	The user credentials are stored ,which is used to send notification of any updates	MySQL
5.	Cloud Database	Database Service on Cloud	IBM DB2, IBM Cloudant etc.
6.	File Storage	File storage requirements	IBM Block Storage or Other Storage Service or Local Filesystem
7.	Machine Learning Model	The model is used to predict whether the student is eligible or not.	Object Recognition Model, etc.
8.	Infrastructure (Server / Cloud)	Application Deployment on Local System / Cloud Local Server Configuration: Cloud Server Configuration :	Local, Cloud Foundry, Kubernetes, etc.

S.No	Characteristics	Description	Technology
1.	Open-Source Frameworks	To create an user friendly interface and to route the data to machine learning model	Flask
2.	Security Implementations	Authorization access scenarios and definitions, hand-over procedures for patient records between wards	IBM Watson STT service
3.	Scalable Architecture	Horizontal scaling is provided by adding more machines to the pool of servers. Vertical scaling is achieved by adding more CPU and RAM to the existing machines.	IBM Watson STT service
4.	Availability	The web dashboard must be available to US and IND users 99.98 percent of the time every month during business hours EST & IST.	IBM cloud and browsers
5.	Performance	The landing page supporting 5,000 users per hour must provide 6 second or less response time in a Chrome desktop browser, including the rendering of text and images and over an LTE connection.	APM technology

Table 3.2.1 Components & Technologies

HTML,CSS & JAVASCRIPT:

HTML (Hypertext Markup Language) is a text-based approach to describing how content contained within an HTML file is structured. This markup tells a web browser how to display text, images and other forms of multimedia on a webpage. HTML is a formal recommendation by the World Wide Web Consortium (W3C) and is generally adhered to by all major web browsers, including both desktop and mobile web browsers. HTML5 is the latest version of the specification.HTML is a text file containing specific syntax, file and naming conventions that show the computer and the web server that it is in HTML and should be read as such.

Cascading Style Sheets (CSS) is a stylesheet language used to describe the presentation of a document written in HTML or XML . CSS describes how elements should be rendered on screen, on paper, in speech, or on other media.CSS is among the core languages of the open web and is standardized across Web browsers according to W3C specifications. Previously, the development of various parts of CSS specification was done synchronously, which allowed the versioning of the latest recommendations.CSS also has rules for alternate formatting if the content is accessed on a mobile device.

JavaScript is a lightweight, interpreted programming language. It is designed for creating network-centric applications. JavaScript is very easy to implement because it is integrated with HTML. It is open and cross-platform.Javascript helps you create really beautiful and crazy fast websites. You can develop your website with a console like look and feel and give your users the best Graphical User Experience.JavaScript simply adds dynamic content to websites to make them look good and HTML work on the look of the website without the interactive effects and all.JavaScript manipulates the content to create dynamic web pages whereas HTML pages are static which means the content cannot be changed.JavaScript is not cross-browser compatible whereas HTML is cross-browser compatible.JavaScript can be embedded inside HTML but HTML can not be embedded inside JavaScript.

PYTHON:

Python is a very popular general-purpose interpreted, interactive, object-oriented, and high-level programming language. Python is dynamically-typed and garbage-collected programming language.Python supports multiple programming paradigms, including Procedural, Object Oriented and Functional programming language. Python design philosophy emphasizes code readability with the use of significant indentation.It supports functional and structured programming methods as well as OOP.It can be used as a scripting language or can be compiled to byte-code for building large applications.It provides very high-level dynamic data types and supports dynamic type checking.It supports automatic garbage collection.It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.

PyCharm:

PyCharm is an Integrated Development Environment (IDE) used for programming in Python. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django. PyCharm is developed by the Czech company JetBrains.It is cross-platform working on Windows, Mac OS X and Linux. PyCharm has a Professional Edition, released under a proprietary license and a Community Edition released under the Apache License.

JUPITER NOTEBOOK:

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. Its uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.Jupyter Notebook (formerly IPython Notebooks) is a web-based interactive computational environment for creating Jupyter notebook documents. The “notebook” term can colloquially make reference to many different entities, mainly the Jupyter web application, Jupyter Python web server, or Jupyter document format depending on context.According to the official website of Jupyter, Project Jupyter exists to develop open-source software, open-standards, and services for interactive computing across dozens of programming languages. Jupyter Book is an open-source project for building books and documents from computational material. It allows the user to construct the content in a mixture of Markdown, an extended version of Markdown called MyST, Maths & Equations using MathJax, Jupyter Notebooks, reStructuredText, the output of running Jupyter Notebooks at build time. Multiple output formats can be produced (currently single files, multipage HTML web pages and PDF files).

FLASK:

Flask is a web application framework written in Python. It was developed by Armin Ronacher, who led a team of international Python enthusiasts called Poocco. Flask is based on the Werkzeg WSGI toolkit and the Jinja2 template engine.Both are Pocco projects.The Web Server Gateway Interface (Web Server Gateway Interface, WSGI) has been used as a standard for Python web application development. WSGI is the specification of a common interface between web servers and web applications.Werkzeug is a WSGI toolkit that implements requests, response objects, and utility functions. This enables a web frame to be built on it. The Flask framework uses Werkzeg as one of its bases.Jinja2 is a popular template engine for Python.A web template system combines a template with a specific data source to render a dynamic web page.

TENSORFLOW:

TensorFlow is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries, and community resources that lets researchers push the state-of-the-art in ML, and gives developers the ability to easily build and deploy ML-powered applications. TensorFlow provides a collection of workflows with intuitive, high-level APIs for both beginners and experts to create machine learning models in numerous languages. Developers have the option to deploy models on a number of platforms such as on servers, in the cloud, on mobile and edge devices, in browsers, and on many other JavaScript platforms. This enables developers to go from model building and training to deployment much more easily.TensorFlow allows developers to create dataflow graphs—structures that describe how data moves through a graph, or a series of processing nodes. Each node in the graph represents a mathematical operation, and each connection or edge between nodes is a multidimensional data array, or *tensor*. If you use Google's own cloud, you can run TensorFlow on Google's custom TensorFlow Processing Unit (TPU) silicon for further acceleration. The resulting models created by TensorFlow, can be deployed on most any device where they will be used to serve predictions.

3.3 SOLUTION & WORKING PROCESS

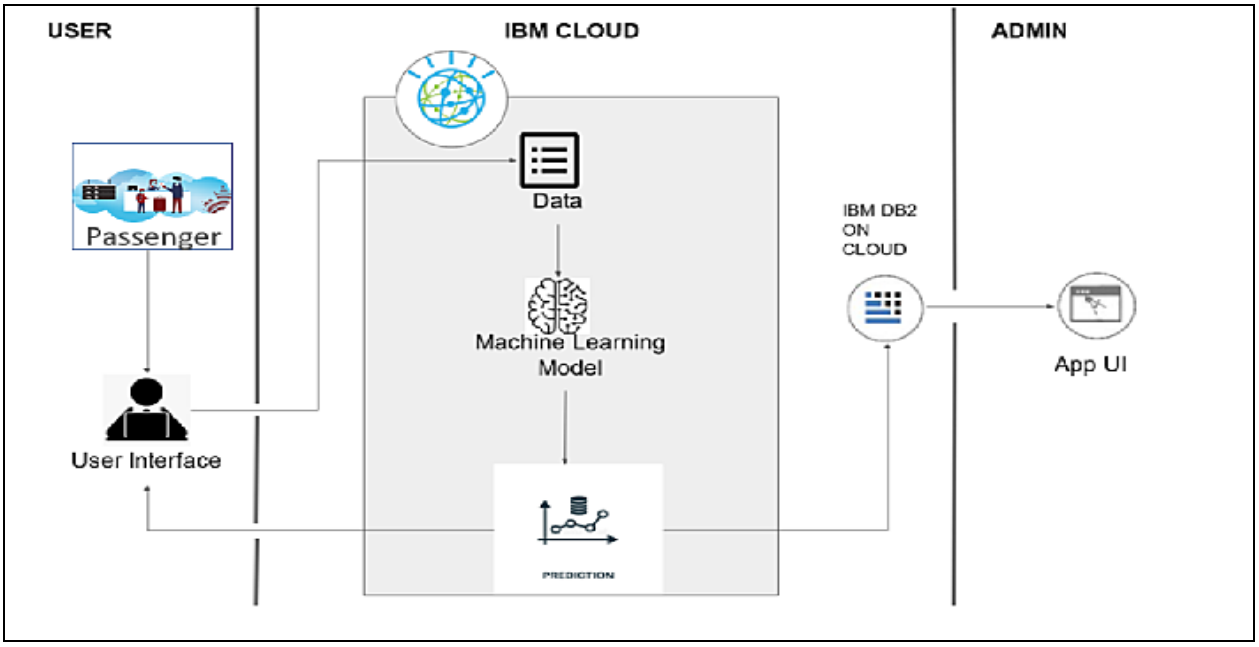


Fig 3.3.1. Technical Architecture

An crucial factor in the proper and timely operation of aircraft is the weather. We provide a technique for anticipating aircraft delays that mostly relies on the current weather conditions. The system needs to be more scalable, so it's important to pick an algorithm that treats each parameter as an independent variable. As the name suggests, supervised learning involves a supervisor serving as an instructor. Essentially, supervised learning is a type of learning where we train or educate the computer using data that has already been correctly labelled, or data that has been appropriately tagged.

After that, the machine is given a fresh collection of examples (data), and an algorithmic procedure for supervised learning analyses the coaching knowledge (set of training instances) and generates an accurate result from tagged data. The labelled data offers the supervised machine learning method's results authenticity. The naive bayed model is one of the algorithms that has been shown to be effective for real-time prediction, and the fact that it treats each characteristic as independent of the others makes it a suitable method for the project in question.

In this study, a final dataset is created by combining three independent data sets. Using the flight data's tail number column and the aircraft data's nnumber column, which corresponds to the tail numbers of the planes, the plane data is combined with the flight data. Since some of the tail number rows originally began with "N," "N" values are eliminated using a lambda function that performs a specified operation on all rows. Both datasets make use of the distinct carrier id, origin airport id, and destination airport id columns to be able to contain the passenger data. Values with a distance greater than 1.5 IQR are detected as outliers and eliminated during the outlier treatment stage.To take advantage of potential insights, the data is used to generate three new features.It is believed that the probability of a delay will increase when a plane operates more trips in a day. Daily flight order is the first feature that has been developed. The data is initially sorted by the day of the month, the tail number, and the CRS departure time before being used to generate this feature. Following this ordering, the first flight is given the number 1 as the daily flight order, and this number is then increased by 1 for each subsequent flight by the CRS departure time, which has the same day of the month and tail number information. It is anticipated that as a carrier operates more flights from an airport at hourly intervals, the workload may raise the probability of delays.

The data is sorted by day of the month, unique carrier, origin airport ID, and departure time block to produce the second characteristic, which is the carrier flight order. Following this ordering, the first flight is given the number 1 as the carrier flight order, and this number is then increased by 1 for each subsequent flight by the departure time block that has the same day of the month, a different carrier, and information about the origin airport ID.

The data is sorted according to the day of the month, the ID of the origin airport, and the departure time interval to produce the airport flight order.

Following this ordering, the first flight is given the number 1 as the airport flight order, and this number is then increased by 1 for each subsequent flight by the departure time block with the same day of the month and origin airport id information with the hope that airport traffic will be taken into account.

3.4 FUNCTIONAL REQUIREMENTS

The following table describes the functional requirements

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration through Form Registration through Gmail Registration through LinkedIN
FR-2	User Confirmation	Confirmation via Email Confirmation via OTP
FR-3	Flight data export process	A Process where web client passenger take flight data from web server to be analyzed in web client
FR-4	Flight Data Management	Web portal admin is able to add, edit, or erase passenger's data
FR-5	Data Management	Web Portal admin is able to add, edit, or erase department data
FR-6	Data Retention	• Proposed application system handle archival, retrieval, and retention of historical data. • Provide sufficient details concerning these.
FR-7	Compliance requirements	This paper proposes a model for predicting flight delay based on Decision Tree. Decision tree is one of the newest methods employed in solving problems with high level of complexity and massive amount of data.

Table 3.4.1 Functional Requirements

The following tables describe the non-functional requirements:

FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	Being able to predict the delay allows for better operational planning at the destination airport based on expected flight delay at origin.
NFR-2	**Security**	We aim to integrate multiple data source to predict the departure delay of scheduled flight.
NFR-3	**Reliability**	Considering the long short- term memory network is prone to over-fitting in limited data sets
NFR-4	**Performance**	• That has been performed based on statistics delay time has been considered to be reduced.
		• It has predicted the delay at destination based on factors that occur in the vicinity of arrival time at destination.
NFR-5	**Availability**	The web dashboard must be available to US and IND users 99.98 percent of the time every month during business hours EST & IST.
NFR-6	**Scalability**	• Flight delays a primary issue for airlines and travelers. • The predicted arrival delay takes into consideration both flight information and weather conditions at origin airport and destination airport according to the flight timetable.

Table 3.4.2 Non-Functional Requirements

3.5 FLOW CHART

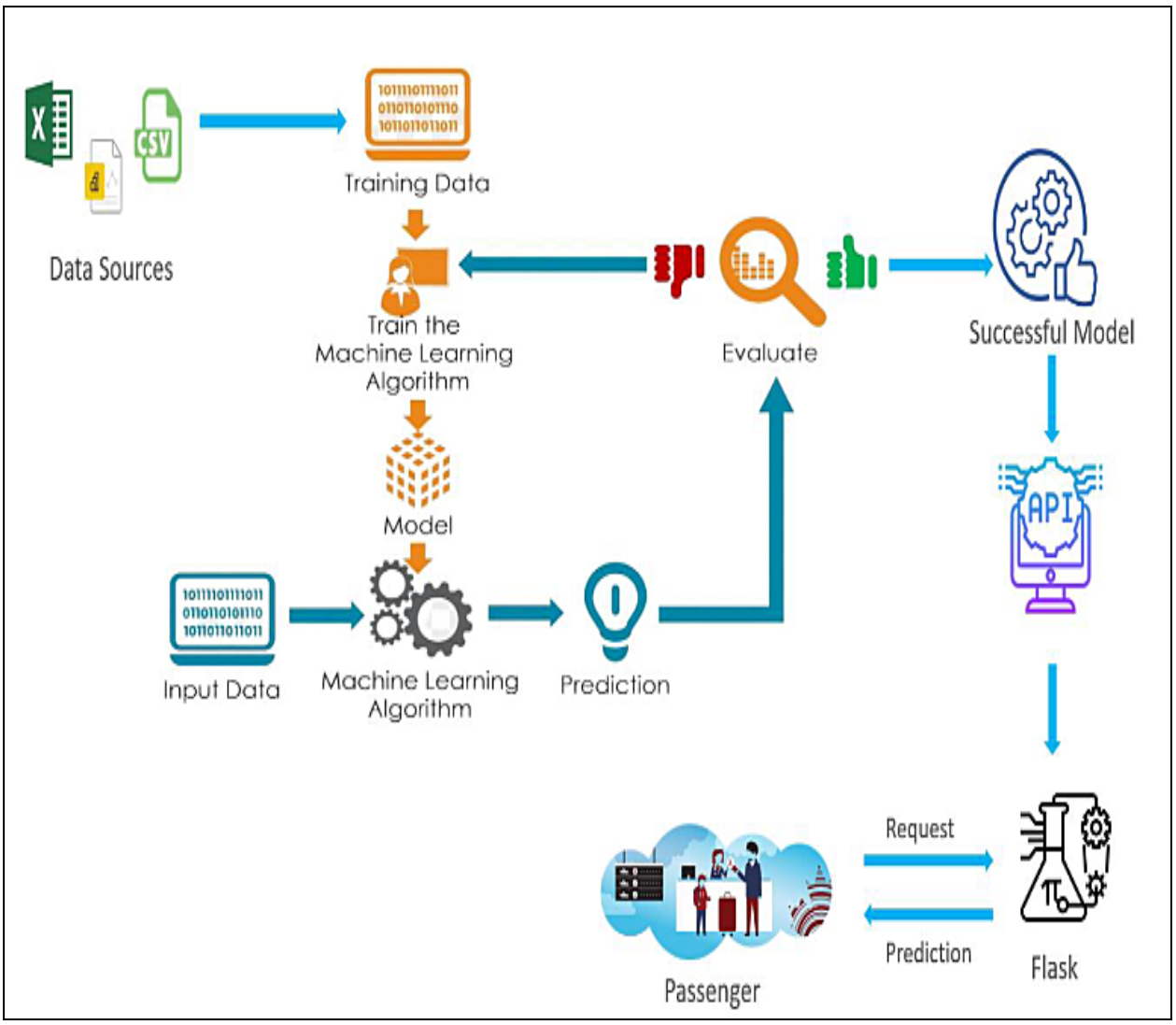
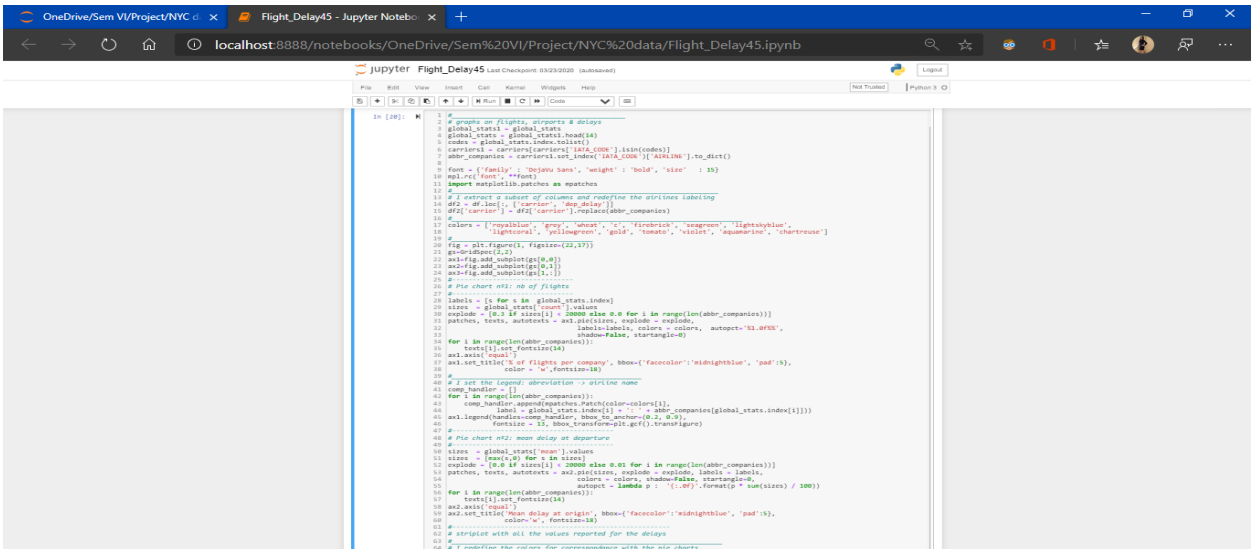
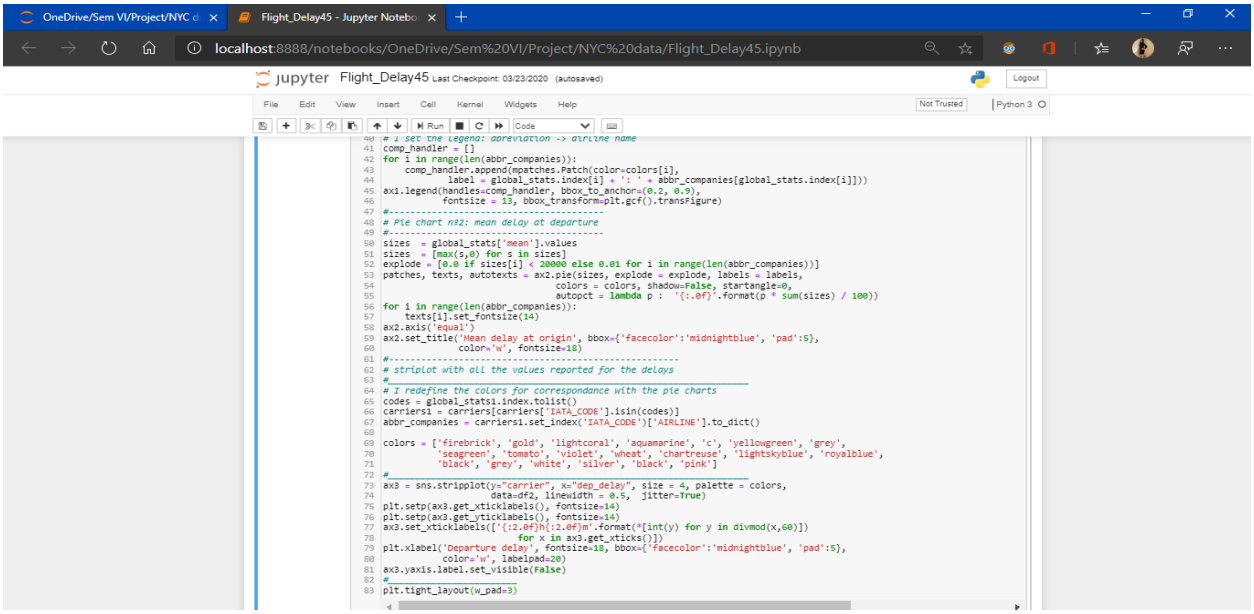


Fig 3.5.1 Solution Architecture

CHAPTER 4
RESULT & SCREEN SHOTS

We found the following factors to be significant after applying both the models for predicting whether a flight should be delayed and how much one would anticipate a flight to be delayed: week, month, airline carrier reference, planned elapsed time (in airtime), distance between two departure and destinations, flight planned departure time, departure airport code, and taxi-in and taxi-out4 time. One may estimate whether a flight might be delayed and, more crucially, how long of a delay one might anticipate by applying our model to the data gathered.





OneDrive/Sem VI/Project/NYC d | Flight_Delay45 - Jupyter Notebo

localhost:8888/notebooks/OneDrive/Sem%20VI/Project/NYC%20data/Flight_Delay45.ipynb

jupyter Flight_Delay45 Last Checkpoint: 03/23/2020 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3

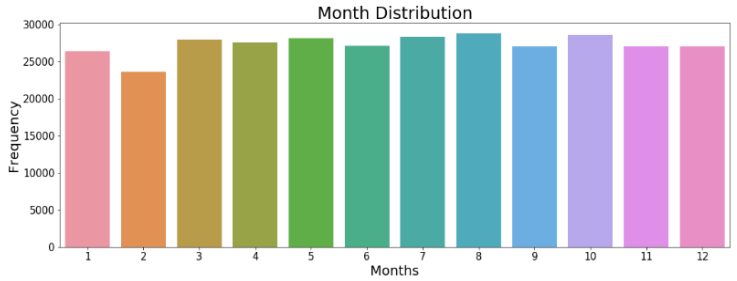
Run Code

Exploratory Data Analysis

Let's check the number of flights per `month`

In [16]:

```
plt.figure(figsize = (18, 6))
sns.countplot(df['month'])
plt.title('Month Distribution', size = 25)
plt.xticks(size = 15)
plt.yticks(size = 15)
plt.xlabel("Months", size = 20)
plt.ylabel("Frequency", size = 20)
plt.show()
```



Flights are distributed fairly evenly among the months, but there is a slight decline at the beginning of the year in the month of **January** and **February**

OneDrive/Sem VI/Project/NYC d | Flight_Delay45 - Jupyter Notebo

localhost:8888/notebooks/OneDrive/Sem%20VI/Project/NYC%20data/Flight_Delay45.ipynb

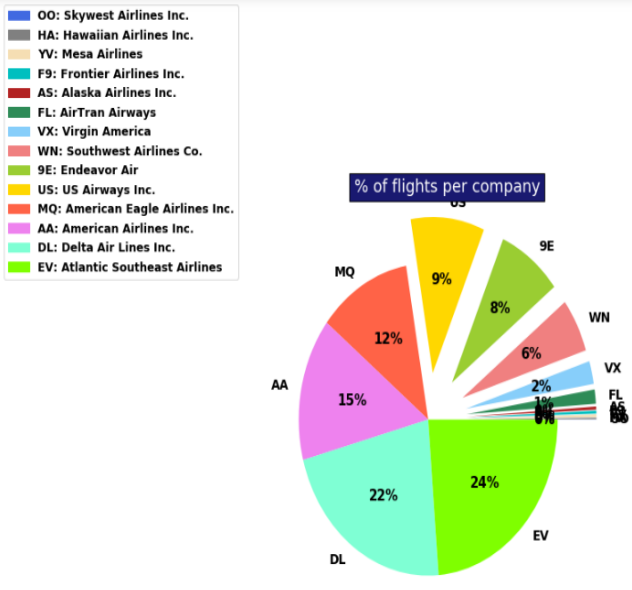
jupyter Flight_Delay45 Last Checkpoint: 03/23/2020 (autosaved)

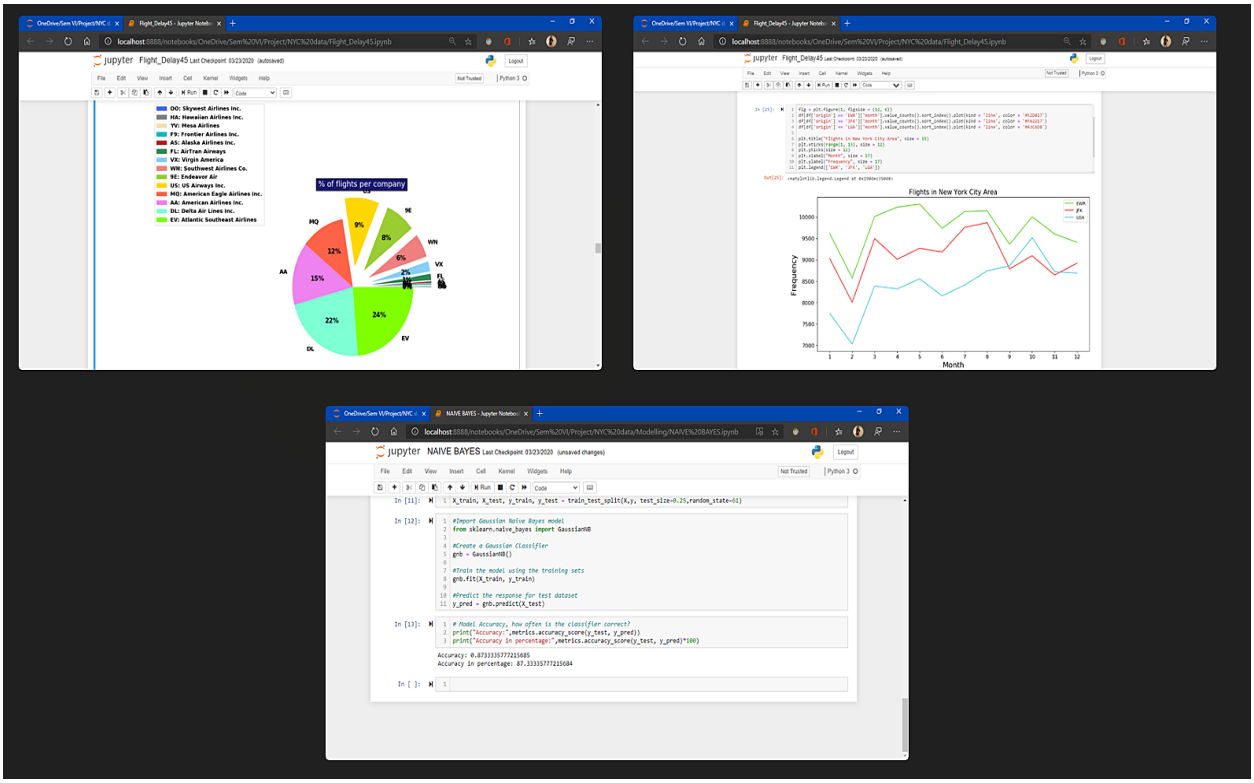
Logout

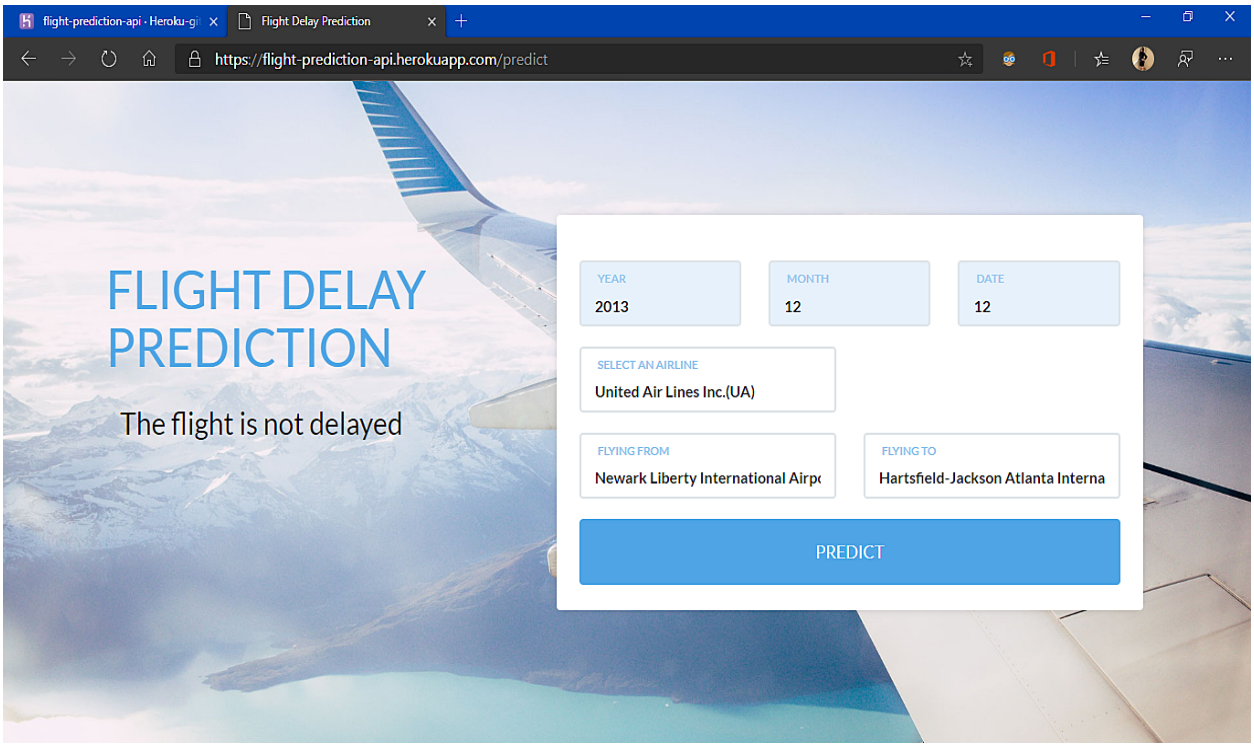
File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3

Run Code




% of flights per company
Flights in New York City Area
Frequency
Month
NAIVE BAYES
Accuracy: 0.8733335777215685
Accuracy in percentage: 87.33335777215684


flight-prediction-api - Heroku-g
Flight Delay Prediction
https://flight-prediction-api.herokuapp.com/predict
FLIGHT DELAY PREDICTION
The flight is not delayed
YEAR
2013
MONTH
12
DATE
12
SELECT AN AIRLINE
United Air Lines Inc.(UA)
FLYING FROM
Newark Liberty International Airpo
FLYING TO
Hartsfield-Jackson Atlanta Interna
PREDICT

CHAPTER 5
CONCLUSION & FUTURE SCOPE

Therefore, to create a better model, researchers should strive to gather more relevant data and use more powerful computing resources. This study explored supervised learning techniques to propose a system for forecasting total flight departure delays in airports. By doing so, we may be able to forecast flight delays without the requirement for several months of data to do so. A further development would be to expand the model's application to include all flights worldwide, or at the very least, to use other data sources to create predictions that are more thorough.The most intriguing step would be to incorporate such a model into a flight booking tool so that future travellers might receive the delay estimate. Even this would require a high level of confidence in the information provided, taking into account the potential impact on reservations.

Classification or reggrerssion ways are often accustomed determine the delay which includes Feed forward network, Neural Network, Random Forrest, decision tress, Naïve Bayes Classification Tree, Regression Tree, etc. As seen from the articles and papers these methodologies offer virtually identical accuracy however we want an algorithmic rule that is good with real world prediction and analysis and thus: naïveBayes. except being smart with real time prediction algorithmic rule that considers or assumes independence among predictors that makes the system scalable as other independent attribute may be superimposed up to the algorithmic rule for computation of the delay. the expected delay can thus facilitate the ground employees for creating correct and smooth operation plans and therefore the data if sent to the passengers will profit the airlines also because the passengers

REFERENCES

[1] Bureau of Transportation Statistics. (2018). On-Time: Marketing Carrier On-Time Performance Retrieved from https://www.transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline%20OnTime%20Performance%20Data&DB_Short_Name=On-Time

[2] Bureau of Transportation Statistics. (2018). T-100 Domestic Segment (U.S. Carriers) Retrieved from https://www.transtats.bts.gov/Tables.asp?DB_ID=110&DB_Name=Air%20Carrier%20Statistics%20%28Form%2041%20Traffic%29-%20%20U.S.%20Carriers&DB_Short_Name=Air%20Carriers#

[3] Federal Aviation Administration. (2019). Releasable Aircraft Database Retrieved from https://www.faa.gov/licenses_certificates/aircraft_certification/aircraft_registry/releasable_aircraft_download/

[4] Elliot, A. F. (2019). How long does it take to turn a plane around – and what's the fastest way to board? The Telegraph. Retrieved from https://www.telegraph.co.uk/travel/travel-truths/plane-turnaround-procedures/

[5] Barro, J. (2019). Here's Why Airplane Boarding Got So Ridiculous. The New York Magazine Inelligencer Retrieved from http://nymag.com/intelligencer/2019/05/heres-why-airplane-boarding-got-soridiculous.html

[6] Chen, Stephen (2019). Failing to Land Flight Delay Predictions https://towardsdatascience.com/failing-to-land-flight-delay-predictions-a281689dd602

[7] Holden, Richard (2018). For the holidays and beyond, your travel planning guide is here https://www.blog.google/products/flights-hotels/travel-planning-guide-for-theholidays-and-beyond/

[8] Martinez Vincent (2012). Flight Delay Prediction (Master's Thesis). Available from https://www.semanticscholar.org/paper/Flight-Delay-Prediction-Master-ThesisMartinez/84daebe3efe5cdc9678d791d4897cd16fc197d92 16

[9] Sternberg, Alice & Soares, Jorge & Carvalho, Diego & Ogasawara, Eduardo. (2017). A Review on Flight Delay Prediction.

[10] Zonglei, Lu & Jiandong, Wang & Guansheng, Zheng. (2009). A New Method to Alarm Large Scale of Flights Delay Based on Machine Learning. 589-592. 10.1109/KAM.2008.18

[11] Khaksar, Hassan & Sheikholeslami, Abdolrreza(n.d.). Airline Delay Prediction by Machine Learning Algorithms.
http://scientiairanica.sharif.edu/article_20020_d43b2e5c29cb07fcb651da6bd2005d30.pdf

[12] Mohtadi Ben Fraj (2017 December 24). In-Depth: Parameter tuning for Gradient Boosting https://medium.com/all-things-ai/in-depth-parameter-tuning-for-gradient-boosting3363992e9bae

[13] Avinash, Navlani (2018 December 28). Decision Tree Classification in Python. Retrieved from https://www.datacamp.com/community/tutorials/decision-tree-classification-python

[14] Breiman, L. Machine Learning (2001) 45: 5. https://doi.org/10.1023/A:1010933404324

[15] Dan, Nelson (2019 July 17) Gradient Boosting Classifiers in Python with Scikit-Learn Retrieved from https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/