**CS 451 – Data Mining**

**Project Report**

**Project – Flight Delay Prediction using Scalable Data Mining**

**Team Members –**

Aakash Sinha – U101114FCS214

Rohan Shrivastava – U101114FCS118

Krishnapriya – U101114FCS189

**Course Mentor –**

Ms. Pragya Verma

**Project Repository Link on Github –**

<https://github.com/aakashsinha19/Flight-Delay-Prediction-Using-Scalable-Data-Mining>

**Work Flow -**

|  |  |  |
| --- | --- | --- |
| PHASE-I | |  | | --- | | Choosing a suitable Dataset  • Understanding the data  • Data Visualization | |
| +PHASE-II | |  | | --- | | • Attribute Selection  • Data Reduction  • Replacing Missing Values  • Removing Outliers  Binning by Mean | |
| PHASE-III | |  | | --- | | Creating Training and Testing Dataset  • Applying different Models like Naïve Bayes, KNN, Decision Trees  • Evaluation of all the algorithms  • Deployment  • Testing  • Data Visualization | |

**INTRODUCTION –**

Delay is one of the most remembered performance indicator of any transportation system. Notably, commercial aviation players understand delay as the period of time by which a flight is late or postponed. Thus, delay may be represented by the difference between scheduled and real times of departure or arrival of a flight. Country regulator authorities have a multitude of indicators related to tolerance thresholds for flight delays. Indeed, flight delay is an important subject in the context of air transportation systems, which indicates how relevant this indicator is and how it affects no matter the scale of airline meshes.

Flight delays have a negative effect on airlines, airports and passengers. Their prediction is crucial during the decision-making process for all players of commercial aviation. The development of accurate prediction models for flight delays became cumbersome due to the complexity of air transportation system, the amount of methods for prediction, and the deluge of data related to such system.

**ABSTRACT**

Flight delays have a negative effect on airlines, airports and passengers. Their prediction is crucial during the decision-making process for all players of commercial aviation. Moreover, the development of accurate prediction models for flight delays became cumbersome due to the complexity of air transportation system, the amount of methods for prediction, and the deluge of data related to such system.

In this context, this paper presents a thorough literature review of approaches used to build flight delay prediction models from the Data Science perspective.

We propose taxonomy and summarize the initiatives used to address the flight delay prediction problem, according to scope, data and computational methods, giving special attention to an increasing usage of machine learning methods. Besides, we also present a timeline of major works that depicts relationships between flight delay prediction problems and research trends to address them.

**Motivation**

1. A flight is on-time if the departure delay is within 15 minutes of the scheduled departure time
2. A flight is delayed if the departure delay is more than 15 minutes late from the scheduled departure time

**PHASE-I**

1. **Data Collection –**

* We have used Kaggle Dataset for Flight Delay and Cancellation 2015 by Department of Transportation.
* The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report and in this dataset of 2015 flight delays and cancellations.
* The flight delay and cancellation data was collected and published by the DOT's Bureau of Transportation Statistics.
* The accuracy of data was 98.5%

1. **Understanding the Data**
2. **Data Visualization –**

* Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.
* This method is being used in the initial stage of our project to analyze the data and find some patterns or inaccuracy if any.
* We are using Microsoft Power BI and Excel.
* **Microsoft Power BI** is a free, self-service business intelligence cloud service that provides non-technical business users with tools for aggregating, analyzing, visualizing and sharing data
* Microsoft **Excel** is a spreadsheet developed by Microsoft for Windows, macOS, Android and iOS. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications

**PHASE- II**

1. **Data Sanitization –**

**Data Sanitization** is the process of making sensitive information in non-production databases safe for wider visibility. This White Paper is an overview of various techniques which can be used to sanitize sensitive production data in test and development databases.

1. **Attribute Selection –**

In [machine learning](https://en.wikipedia.org/wiki/Machine_learning) and [statistics](https://en.wikipedia.org/wiki/Statistics), feature selection, also known as variable selection, **attribute selection** or variable subset selection, is the process of selecting a subset of relevant [features](https://en.wikipedia.org/wiki/Feature_(machine_learning)) (variables, predictors) for use in model construction. Feature selection techniques are used for four reasons:

* simplification of models to make them easier to interpret by researchers/users,
* shorter training times,
* to avoid the [curse of dimensionality](https://en.wikipedia.org/wiki/Curse_of_dimensionality),
* Enhanced generalization by reducing [over fitting](https://en.wikipedia.org/wiki/Overfitting) (formally, reduction of [variance](https://en.wikipedia.org/wiki/Bias-variance_tradeoff)).

1. **Attributes –**

**i. Airlines**

UA - United Air Lines Inc. (Selected)

AA - American Airlines Inc. (Selected)

US - US Airways Inc. (Selected)

F9 - Frontier Airlines Inc.

B6 - JetBlue Airways

OO - Skywest Airlines Inc.

AS - Alaska Airlines Inc.

NK - Spirit Air Lines

WN - Southwest Airlines Co.

DL - Delta Air Lines Inc. (Selected)

EV - Atlantic Southeast Airlines (Selected)

HA - Hawaiian Airlines Inc. (Selected)

MQ - American Eagle Airlines Inc. (Selected)

VX - Virgin America (Selected)

**ii. Airports (Selected)**

ABY - Southwest Georgia Regional Airport

ACY - Atlantic City International Airport

ALO - Waterloo Regional Airport

BHM - Birmingham-Shuttlesworth International Airport

BWI - Baltimore-Washington International Airport

COU - Columbia Regional Airport

DTW - Columbia Regional Airport

CVG - Cincinnati/Northern Kentucky International Airport

CDC - Cedar City Regional Airport

BTV - Burlington International Airport

1. **Data Subset Selection –**

Selecting is one of the most used and important operation for manipulating, editing, organizing content in a document. By selecting elements in document it enables you to copy, move, and delete them or to edit their properties.

1. **Replacing Missing Values –**

Missing observations can be problematic in analysis, and some time series measures cannot be computed if there are missing values in the series. Sometimes the value for a particular observation is simply not known. In addition, missing data can result from any of the following:

* + - Each degree of differencing reduces the length of a series by 1.
    - Each degree of seasonal differencing reduces the length of a series by one season.
    - If you create new series that contain forecasts beyond the end of the existing series (by clicking a **Save** button and making suitable choices), the original series and the generated residual series will have missing data for the new observations.
    - Some transformations (for example, the log transformation) produce missing data for certain values of the original series. Missing data at the beginning or end of a series pose no particular problem; they simply shorten the useful length of the series. Gaps in the middle of a series (*embedded* missing data) can be a much more serious problem. The extent of the problem depends on the analytical procedure you are using.

**Binning by Mean –**

* We have used Binning by Means.
* Binning methods smooth a sorted data value by consulting its “neighborhood”, that is, the values around it.
* The sorted values are distributed into a number of buckets or bins.
* In **smoothing by bin means**, each value in a bin is replaced by the mean value of the bin.
* Similarly, **smoothing by bin medians** can be employed, in which each bin value is replaced by the bin median.
* In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

**Binning Results –**

* Calculated Mean for SCHEDULED\_DEPARTURE = 1319.647527 ~ 1320
* Calculated Mean for ACTUAL\_DEPARTURE = 1336.685433 ~ 1337
* Calculated Mean for DEPARTURE\_DELAY = 12.16711349 ~ 12
* Calculated Mean for SCHEDULED\_ARRIVAL = 1498.20318 ~ 1499
* Calculated Mean for ACTUAL\_ARRIVAL = 1495.914951 ~ 1496
* Calculated Mean for ARRIVAL\_DELAY = 7.961504029 ~ 8

1. **Removing Outliers –**

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In sense, this definition leaves it up to the analyst to decide what will be considered abnormal

.