Flight Delay Prediction Using Machine Learning Algorithm XGBoost

Article *in* Journal of Advanced Research in Dynamical and Control Systems · September 2019

CITATION
1

READS
2,944

2 authors, including:

Subhani Shaik
Sreenidhi Institute of Science & Technology(Autonomous)
57 PUBLICATIONS **56** CITATIONS

SEE PROFILE

Flight Delay Prediction Using Machine Learning Algorithm XGBoost

K.P. Surya Teja, Assistant Professor, Dept. of IT, Sreenidhi Institute of Science and Technology, (Autonomous), Hyderabad. E-mail: kpsuryateja@sreenidhi.edu.in

Vigneswara Reddy, Assistant Professor, Dept. of IT, Sreenidhi Institute of Science and Technology, (Autonomous), Hyderabad. E-mail: vigneswarareddyk@sreenidhi.edu.in

Dr. Shaik Subhani, Associate Professor, Dept. of IT, Sreenidhi Institute of Science and Technology, (Autonomous), Hyderabad. E-mail: shaiksubhani@sreenidhi.edu.in

Abstract--- Growth in aviation industries has resulted in air-traffic jamming causing flight delays. Flight delays not only have economic impact but also injurious environmental properties. Air-traffic supervision is becoming increasingly challenging. Airlines delays make immense loss for business field as well as in budget loss for a country, there are so many reasons for impede in flights some of them are, some of them are due to security issues, mechanical problems, due to weather conditions, Airport congestion etc. we are proposing machine learning algorithms like XGBoost regressed, Linear regression Techniques. The aim of this research work is to predict Flight Delay, Which is highest economy producing field for many countries and among many transportation this one is fastest and comfort, so to identify and reduce flight delays, can dramatically reduce the flight delays to saves huge amount of turnovers, using machine-learning algorithms.

Keywords--- Flight Delay, Linear Regression Techniques, XGBoost Algorithm, Air-traffic.

I. Introduction

As population increases tremendously and time is everything for many billionaire. Here the importance of Flights were raised, but due to high cost and some continuous delay of flight made less eyes on flights in 1960's, but due to government help many companies have been started manufacturing flights with less cost and more comfort and many Airports, this made control of airlines traffic. Airlines Economy play a predominant role in countries economy, so there is huge losses had occurred, we all know recent technology of Machine learning is one of the way to determine the flight delays. Mining techniques for instances applied to airlines topics rise rapidly due to their high concert in predicting outcomes, reducing costs of cancellation, promoting excellent airline transportation, improves customers counting and making real time choice to save people's time, money and completing their work smoothly. Regression is one of the most essential works in machine learning and mining techniques, a lot of research has been conducted to apply mining techniques and machine learning on different data items of Flight delay. Many of them show good and less mean supreme error.

II. Literature Survey

Since two decades, rapid growth in air traffic is observed due to comfort, flexibility, and speed. Every year, huge amount around $22 billion loss is noticed due to delay of flights in U.S as per the reports of FAA (Federal Aviation Administration).According to Federal authorities if delay is more than 3 hours for domestic flights and more than 4 hours for International flights the airlines companies have to pay penalty. To avoid the paying of penalty to customer the airlines companies want to maintain a continues relationship among them. Air transportation provides services in the aviation sector and creates wider socioeconomic settlement through its potential to enable convinced types of actions in a local market. According to U.S taxi-out operations are accountable for 4,000 tons of hydrocarbons, 8,000 tons of nitrogen oxides and 45,000 tons of carbon monoxide emissions in the U.S in 2007. In addition, the economic impact of flight delays for domestic flights in the US is probable to be more than $19 Billion per year to the airlines and over $41 Billion per year to the national economy [3].

III. Dataset Description

To train and test models, we used a publicly available kaggle dataset for United States domestic air traffic. The original source of our dataset is the on-line Bureau and Transportation Statistics database [4].The data set is for the year 2015 and consists of well over 3 Million examples with 19 features categorized as follows:

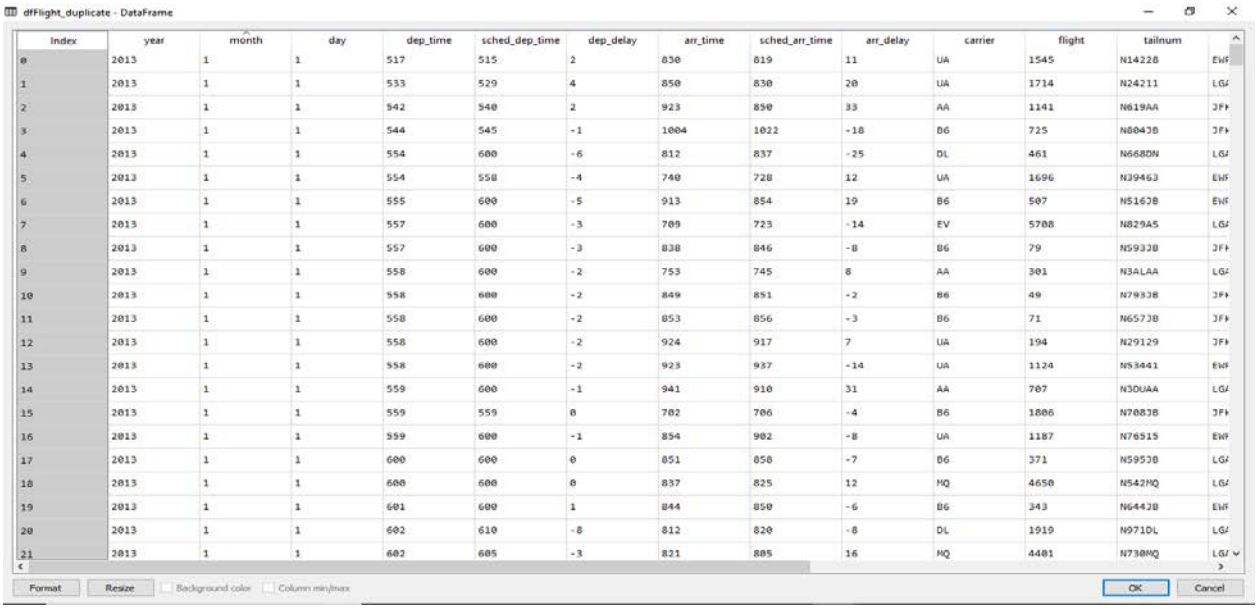
Received: 18 Mar 2019/Accepted: 17 Apr 2019

1. Information about flight (day, month, year, airline, flight number, tail number), 2. Information about origin and destination (origin airport, destination airport), 3. Information about the departure (scheduled departure, departure time, departure delay, tax), 4. Information about the flight-journey (air time, distance, hour, minute, time_hour), 5. Information about the arrival (scheduled arrival, arrival time, arrival delay), 6. Information about tailnum, origin and destination, carrier.

3.1 Loading Dataset

Step 1:- Open the python API module and select the CSV files and browser the data set to a Num_Py array and use it for machine learning. Here, we used read_csv method of pandas to import the dataset. Before improving the dataset, check the present working directory and select the directory where the data is available.

Table 1: Dataset Description


dfFlight_duplicate - DataFrame

Index	year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tailnum	
0	2013	1	1	517	515	2	830	819	11	UA	1545	N14228	EWR
1	2013	1	1	533	529	4	850	830	20	UA	1714	N24211	LGA
2	2013	1	1	542	540	2	923	850	33	AA	1141	N619AA	JFK
3	2013	1	1	544	545	-1	1004	1022	-18	B6	725	N804JB	JFK
4	2013	1	1	554	600	-6	812	837	-25	DL	461	N668DN	LGA
5	2013	1	1	554	558	-4	740	728	12	UA	1696	N39463	EWR
6	2013	1	1	555	600	-5	913	854	19	B6	507	N516JB	EWR
7	2013	1	1	557	600	-3	709	723	-14	EV	5708	N829AS	LGA
8	2013	1	1	557	600	-3	838	846	-8	B6	79	N593JB	JFK
9	2013	1	1	558	600	-2	753	745	8	AA	301	N3ALAA	LGA
10	2013	1	1	558	600	-2	849	851	-2	B6	49	N793JB	JFK
11	2013	1	1	558	600	-2	853	856	-3	B6	71	N657JB	JFK
12	2013	1	1	558	600	-2	924	917	7	UA	194	N29129	JFK
13	2013	1	1	558	600	-2	923	937	-14	UA	1124	N53441	EWR
14	2013	1	1	559	600	-1	941	910	31	AA	707	N3DUAA	LGA
15	2013	1	1	559	559	0	702	706	-4	B6	1806	N708JB	JFK
16	2013	1	1	559	600	-1	854	902	-8	UA	1187	N76515	EWR
17	2013	1	1	600	600	0	851	858	-7	B6	371	N595JB	LGA
18	2013	1	1	600	600	0	837	825	12	MQ	4650	N542MQ	LGA
19	2013	1	1	601	600	1	844	850	-6	B6	343	N644JB	EWR
20	2013	1	1	602	610	-8	812	820	-8	DL	1919	N971DL	LGA
21	2013	1	1	602	605	-3	821	805	16	MQ	4401	N730MQ	LGA

Format Resize Background color Column min/max OK Cancel

3.2 Data Frame

A Data frame is a 2-dimensional data structure, i.e., data is associated with a tabular fashion in rows and columns.

Features of Data Frame:

1. Potentially columns are of different types
2. Size – Mutable.
3. Labeled axes (rows and columns)
4. Can Perform Arithmetic operations on rows and columns.

3.3 Dealing with Missing Values

Datasets may contain the missing values, often encoded as blanks, NaNs or other placeholders. Such datasets however are incompatible with scikit-learn estimators which assume that all values in an array are numerical, and that all have and hold meaning. A basic strategy to use incomplete datasets is to discard entire rows and/or columns containing missing values. However, this comes at the price of losing data which may be valuable (even though incomplete). A better strategy is to impute the missing values, i.e., to infer them from the known part of the data. Here we use Imputer class of sklearn module and the methods used are transform and fit_transform.

3.4 Mean, Median and Mode

Computing the overall mean, median or mode is a very basic imputation method, it is the only tested function that takes no advantage of the time series characteristics or relationship between the variables. It is very fast, but has clear disadvantages. One disadvantage is that mean imputation reduces variance in the dataset.

Received: 18 Mar 2019/Accepted: 17 Apr 2019

3.5 Encoding Categorical Data

Categorical data are variables that contain label values rather than numeric values. The number of possible values is often limited to a fixed set. Categorical variables are often called nominal. Many machine learning algorithm, cannot operate on label data directly. They categorical data must be converted to a variables and output variables to be numeric. If the categorical variable is an output variable, you may also want to convert predictions by the model back into a categorical form in order to present them or use them in some application. We use Category Encoders to improve model performances when you have nominal or ordinal data that may provide value.

1. If the categorical features are ordinal ones, use label encoder
2. If non-ordinal relation, use hot encoder.

3.6 Splitting the Dataset

A machine learning algorithm works in two stages-the testing and training stage. The training dataset (also called training set, learning set, or AI training data) is the initial dataset used to train an algorithm to understand how to apply technologies such as neural networks, to learn and produce complex results. It includes both input data and the corresponding expected output. The purpose of the training dataset is to provide your algorithm with “ground truth” data. The test dataset, however, is used to assess how well your algorithm was trained with the training dataset. You can’t simply reuse the training dataset in the testing stage because the algorithm will already “know” the expected output, which defeats the purpose of testing the algorithm. Here’s a cool flowchart that shows the training process and the different functions of training data and test data.

IV. Results and Analysis

Supervised learning algorithms make predictions based on a set of examples. For instance, historical stock prices can be used to hazard guesses at future prices. Each example used for training is labeled with the value of interest—in this case the stock price. A supervised learning algorithm looks for patterns in those value labels. It can use any information that might be relevant—the day of the week, the season, the company’s financial data, the type of industry, the presence of disruptive geopolitical events—and each algorithm looks for different types of patterns. After the algorithm has found the best pattern it can, it uses that pattern to make predictions for unlabeled testing data—tomorrow’s prices.

1. **Classification:** When the data are being used to predict a category, supervised learning is also called classification. This is the case when assigning an image as a picture of either a 'cat' or a 'dog'. When there are only two choices, it's called two-class or binomial classification. When there are more categories, as when predicting the winner of the NCAA March Madness tournament, this problem is known as multi-class classification.
2. **Regression:** When a value is being predicted, as with stock prices, supervised learning is called regression.
3. **Anomaly detection:** Sometimes the goal is to identify data points that are simply unusual. In fraud detection, for example, any highly unusual credit card spending patterns are suspect. The possible variations are so numerous and the training examples so few, that it's not feasible to learn what fraudulent activity looks like. The approach that anomaly detection takes is to simply learn what normal activity looks like (using history non-fraudulent transactions) and identify anything that is significantly different.

4.1 Considerations when Choosing an Algorithm

1) Accuracy

Getting the most accurate answer possible isn't always necessary. Sometimes an approximation is adequate, depending on what you want to use it for. If that's the case, you may be able to cut your processing time dramatically by sticking with more approximate methods. Another advantage of more approximate methods is that they naturally tend to avoid over fitting.

2) Training time

The number of minutes or hours necessary to train a model varies a great deal between algorithms. Training time is often closely tied to accuracy—one typically accompanies the other. In addition, some algorithms are more sensitive to the number of data points than others. When time is limited it can drive the choice of algorithm, especially when the data set is large.

Received: 18 Mar 2019/Accepted: 17 Apr 2019

3) Linearity

Lots of machine learning algorithms make use of linearity. Linear classification algorithms assume that classes can be separated by a straight line (or its higher-dimensional analog). These include logistic regression and support vector machines (as implemented in Azure Machine Learning). Linear regression algorithms assume that data trends follow a straight line. These assumptions aren't bad for some problems, but on others they bring accuracy down.

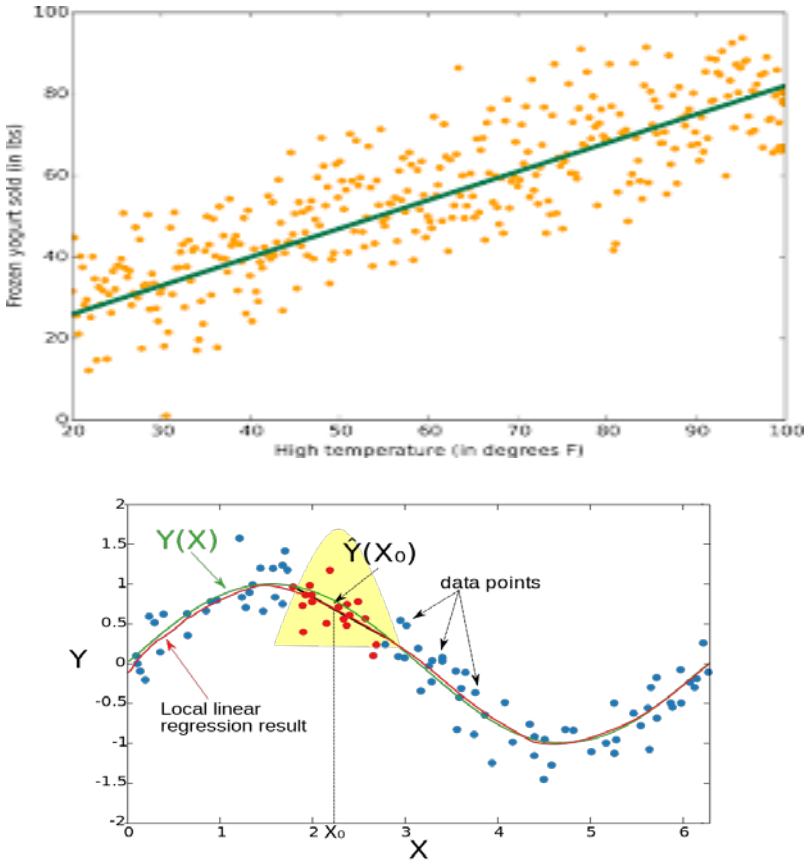


Figure 2: Non-Linear: Trends on Linear Regression is with Low Accuracy

Data with a nonlinear trend - using a linear regression method would generate much larger errors than necessary despite their dangers; linear algorithms are very popular as a first line of attack. They tend to be algorithmically simple and fast to train.

Number of Parameters

Parameters are the knobs a data scientist gets to turn when setting up an algorithm. They are numbers that affect the algorithm's behavior, such as error tolerance or number of iterations, or options between variants of how the algorithm behaves. The training time and accuracy of the algorithm can sometimes be quite sensitive to getting just the right settings. Typically, algorithms with large numbers parameters require the most trial and error to find a good combination.

Alternatively, there is a parameter sweeping module block in Azure Machine Learning that automatically tries all parameter combinations at whatever granularity you choose. While this is a great way to make sure you've spanned the parameter space, the time required to train a model increases exponentially with the number of parameters. The upside is that having many parameters typically indicates that an algorithm has greater flexibility. It can often achieve very good accuracy. Provided you can find the right combination of parameter settings.

Number of Features

For certain types of data, the number of features can be very large compared to the number of data points. This is often the case with genetics or textual data. The large number of features can bog down some learning algorithms, making training time unfeasibly long. Support Vector Machines are particularly well suited to this case.

Received: 18 Mar 2019/Accepted: 17 Apr 2019

Logistic Regression

Logistic regression is another technique borrowed by machine learning from the field of statistics. Logistic Regression, falls under Supervised Machine Learning. It solves the problems of Classification (to make predictions or take decisions based on past data). It is used to predict binary outcomes for a given set of independent variables. The dependent variable's outcome is discrete.

Logistic Function

Logistic regression is named for the function used at the core of the method, the logistic function. The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

$$1 / (1 + e^{-value}) \quad (1)$$

Where e is the base of the natural logarithms (Euler's number or the EXP () function in your spreadsheet) and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function.

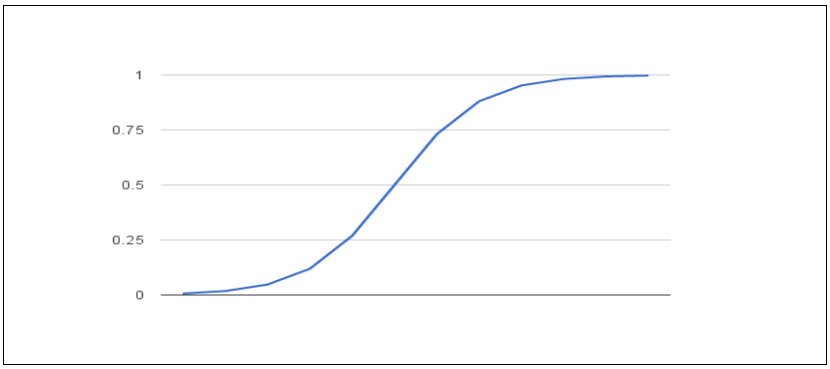


Figure 2: Plot of the Numbers between -5 and 5 Transformed into the Range 0 and 1 Using the Logistic Function

Now that we know what the logistic function is, let's see how it is used in logistic regression.

4.2 Representation Used for Logistic Regression

Logistic regression uses an equation as the representation, very much like linear regression. Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modeled is binary values (0 or 1) rather than a numeric value. Below is an example logistic regression equation:

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)}) \quad (2)$$

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data. The actual representations of the model that you would store in memory or in a file are the coefficients in the equation.

4.3 XGBoost Algorithm

XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. XGBoost is a software library that you can download and install on your machine, then access from a variety of interfaces. Specially, XGBoost supports the following main interfaces:

1. Command Line Interface (CLI).
2. C++ (the language in which the library is written).
3. Python interface as well as a model in scikit-learn
4. R interface as well as a model in the caret package
5. Julia
6. Java and JVM languages like Scala and platforms like Hadoop.

Received: 18 Mar 2019/Accepted: 17 Apr 2019

The library is laser focused on computational speed and model performance; as such there are few frills. Nevertheless, it does offer a number of advanced features. The implementation of the model supports the features of the scikit-learn and R implementations, with new additions like regularization. Three main forms of gradient boosting are supported:

1. Gradient Boosting algorithm also called gradient boosting machine including the learning rate.
2. Stochastic Gradient Boosting with sub-sampling at the row, column and column per split levels
3. Regularized Gradient Boosting with both L1 and L2 regularization.

The implementation of the algorithm was engineered for efficiency of compute time and memory resources. A design goal was to make the best use of available resources to train the model. Some key algorithm implementation features include: Sparse Aware implementation with automatic handling of missing data values, Block Structure to support the parallelization of tree construction, Continued Training so that you can further boost an already fitted model on new data, XGBoost is free open source software available for use under the permissive Apache-2 license. Generally, XGBoost is fast really fast when compared to other implementations of gradient boosting.

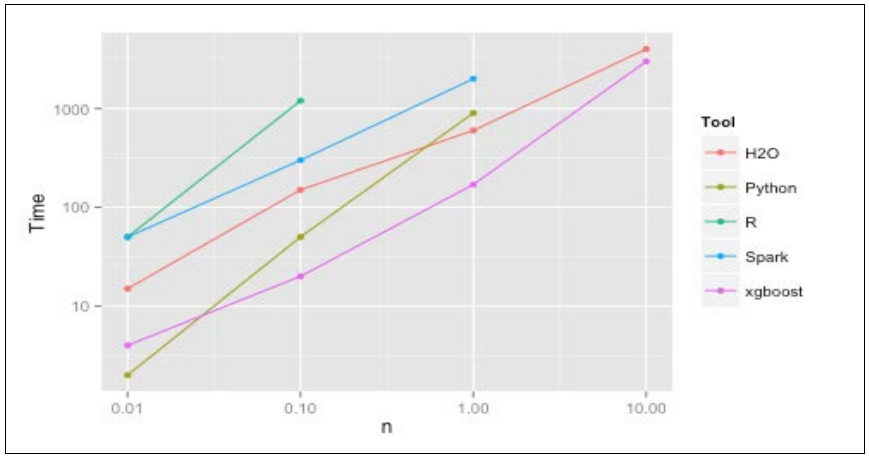


Figure 3: The Implementation of the Algorithm was Engineered for Efficiency of Compute Time and Memory Resources

All scorer objects follow the convention that **higher return values are better than lower return values**. Thus metrics which measure the distance between the model and the data, like metrics.mean_squared_error, are available as neg_mean_squared_error which return the negated value of the metric.

Table 2: Classification Scoring Values

Scoring Classification	**Function**
'accuracy'	metrics.accuracy_score
'percision'	metrics.precision_score
'recall'	metrics.recall_score

Table 3: Regression Scoring Functions

Scoring Classification	**function**
'neg_mean_absolute_error'	metrics.mean_absolute_error
'neg_mean_squared_error'	metrics.mean_squared_error

Our project score using Mean absolute error, In statistics, mean absolute error is a measure of difference between two continuous variables. Assume *X* and *Y* are variables of paired observations that express the same phenomenon. Examples of *Y* versus include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of easurement versus an alternative technique of measurement. Consider a scatter plot of *n* points, where point *I* has coordinates (x_i, y_i)... Mean Absolute Error is the average vertical distance between each point and the identity line. MAE is also the average horizontal distance between each point and the identity line. The Mean Absolute Error is given by:

$$MAE=\sum^{n}_{i=1} | y_i - x_i | /n = \sum^{n}_{i=1} |e_i| /n \quad (3)$$

Received: 18 Mar 2019/Accepted: 17 Apr 2019

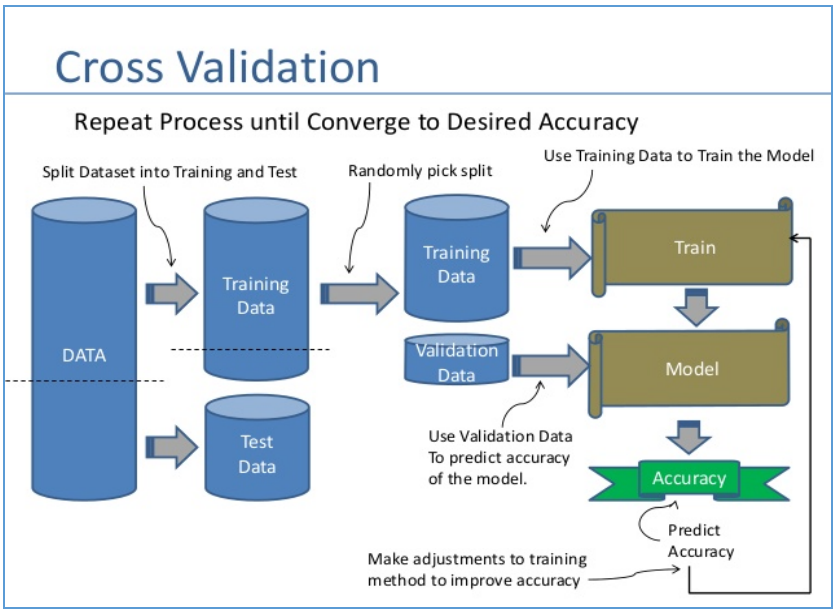


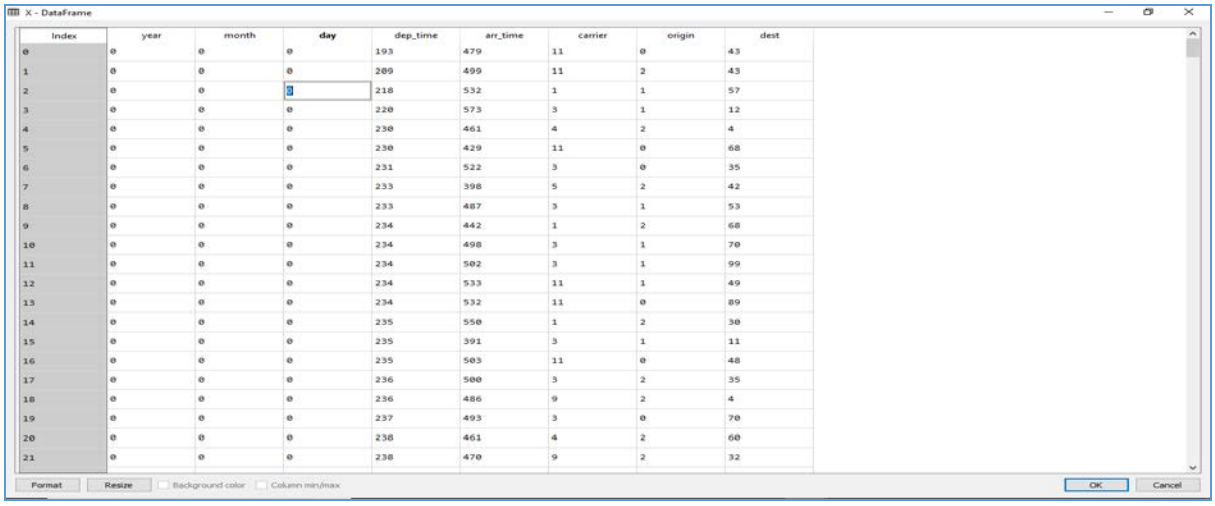
Figure 4: System for Data Crosses Validation

We use train_test_split method of sklearn. Pre-processing module to split the dataset. The dividing ratio has a huge impact on the accuracy of prediction. Usually the training data size should be greater than the testing data size for achieving good accuracy. We have divided the dataset in the ratio 80-20 for good accuracy of the model. Of all the 570 samples of our dataset 456 samples are given to the training set and the remaining 114 samples i.e., 20% of the dataset is given to the testing set.

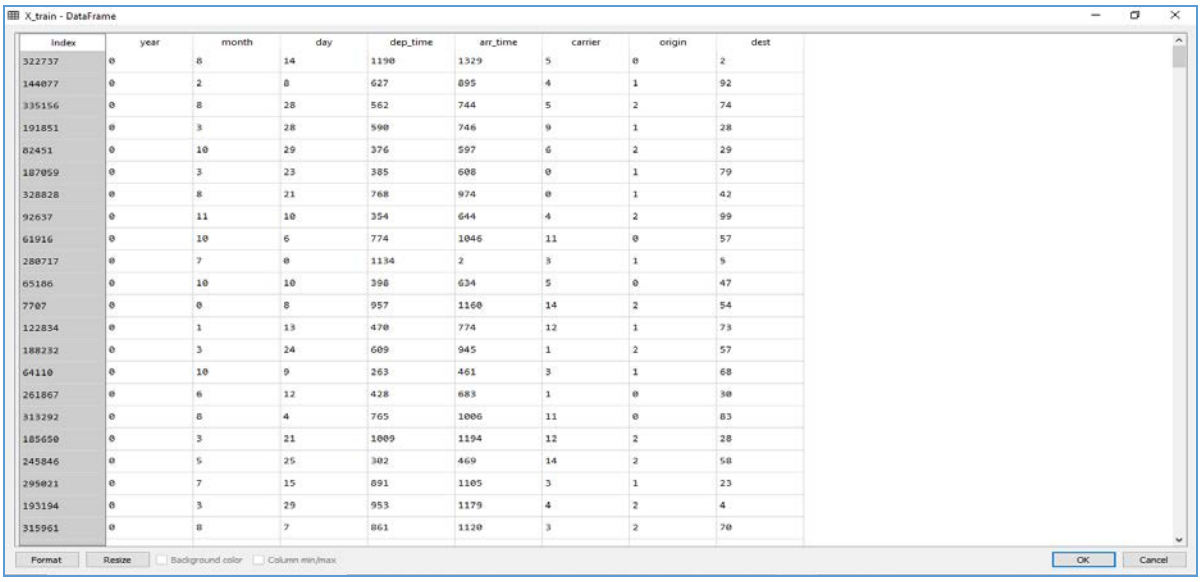
4.4 Data Visualization

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral to yourself and stakeholders than measures of association or significance. The purpose of mesh grid is to create a rectangular grid out of an array of x values and an array of y values. So, for example, if we want to create a grid where we have a point at each integer value between 0 and 4 in both the x and y directions. To create a rectangular grid, we need every combination of the x and y points.

Table 4: A Rectangular Grid Out of an Array of X Values and an Array of Y Values

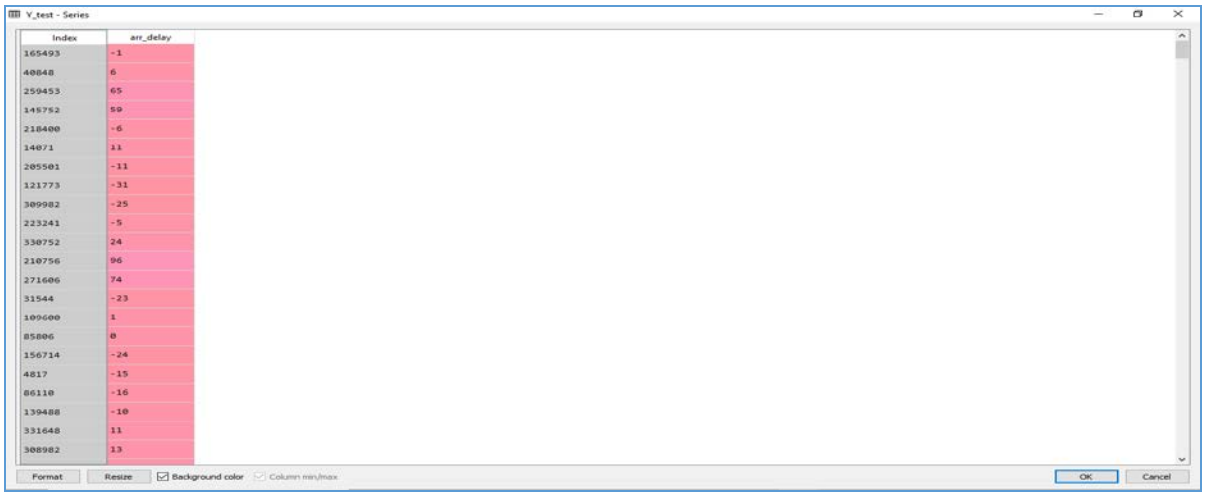


Received: 18 Mar 2019/Accepted: 17 Apr 2019


X_train - DataFrame

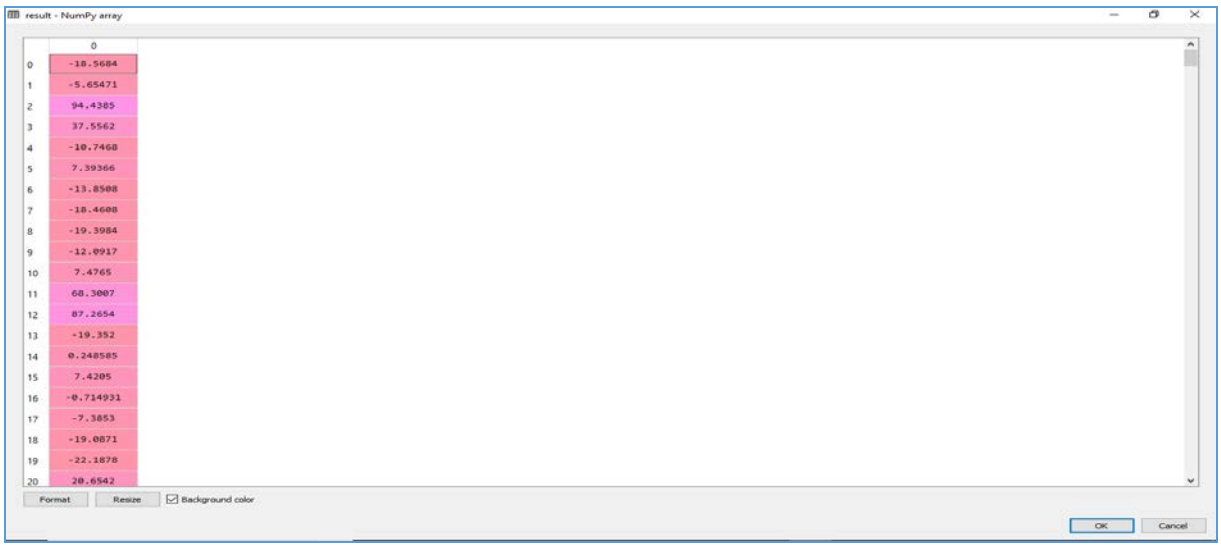
Index	year	month	day	dep_time	arr_time	carrier	origin	dest
322737	0	8	14	1190	1329	5	0	2
144077	0	2	8	627	895	4	1	92
335156	0	8	28	562	744	5	2	74
191851	0	3	28	590	746	9	1	28
82451	0	10	29	376	597	6	2	29
187059	0	3	23	385	608	0	1	79
328828	0	8	21	768	974	0	1	42
92637	0	11	10	354	644	4	2	99
61916	0	10	6	774	1046	11	0	57
280717	0	7	0	1134	2	3	1	5
65186	0	10	10	398	634	5	0	47
7707	0	0	8	957	1160	14	2	54
122834	0	1	13	470	774	12	1	73
188232	0	3	24	609	945	1	2	57
64110	0	10	9	263	461	3	1	68
261867	0	6	12	428	683	1	0	30
313292	0	8	4	765	1006	11	0	83
185650	0	3	21	1009	1194	12	2	28
245846	0	5	25	302	469	14	2	58
295021	0	7	15	891	1105	3	1	23
193194	0	3	29	953	1179	4	2	4
315961	0	8	7	861	1120	3	2	70

Format Resize Background color Column min/max OK Cancel


Y_test - Series

Index	arr_delay
165493	-1
40848	6
259453	65
145752	59
218400	-6
14071	11
205501	-11
121773	-31
309982	-25
223241	-5
330752	24
210756	96
271606	74
31544	-23
109600	1
85806	0
156714	-24
4817	-15
86110	-16
139488	-10
331648	11
308982	13

Format Resize Background color Column min/max OK Cancel


result - NumPy array

	0
0	-18.5684
1	-5.65471
2	94.4385
3	37.5562
4	-10.7468
5	7.39366
6	-13.8508
7	-18.4608
8	-19.3984
9	-12.0917
10	7.4765
11	68.3007
12	87.2654
13	-19.352
14	0.248585
15	7.4205
16	-0.714931
17	-7.3853
18	-19.0871
19	-22.1878
20	20.6542

Format Resize Background color OK Cancel

Top Ten Busiest Flight Routes in order to find traffic highest traffic stations, so that high cancellation may occur at that station.

Received: 18 Mar 2019/Accepted: 17 Apr 2019

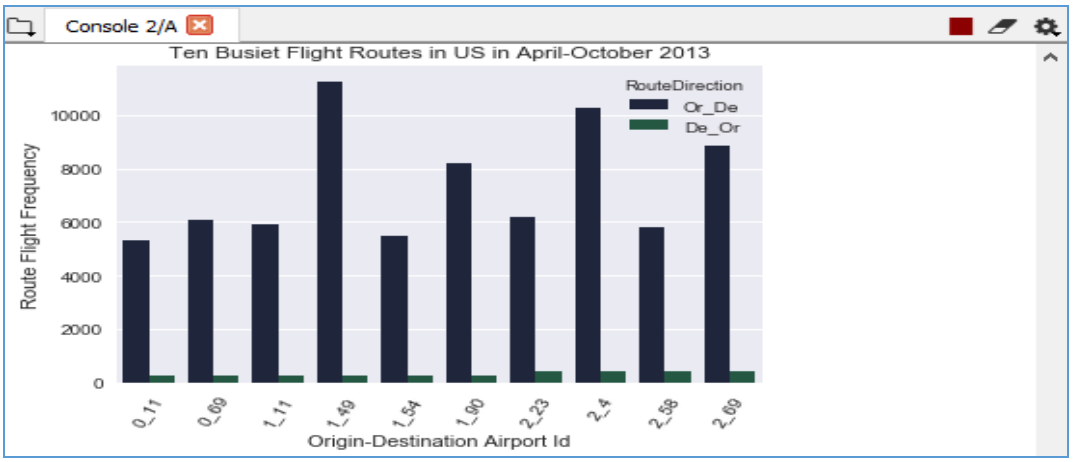


Figure 5: Top Ten Busiest Flight Routes in order to find Traffic Highest Traffic Stations

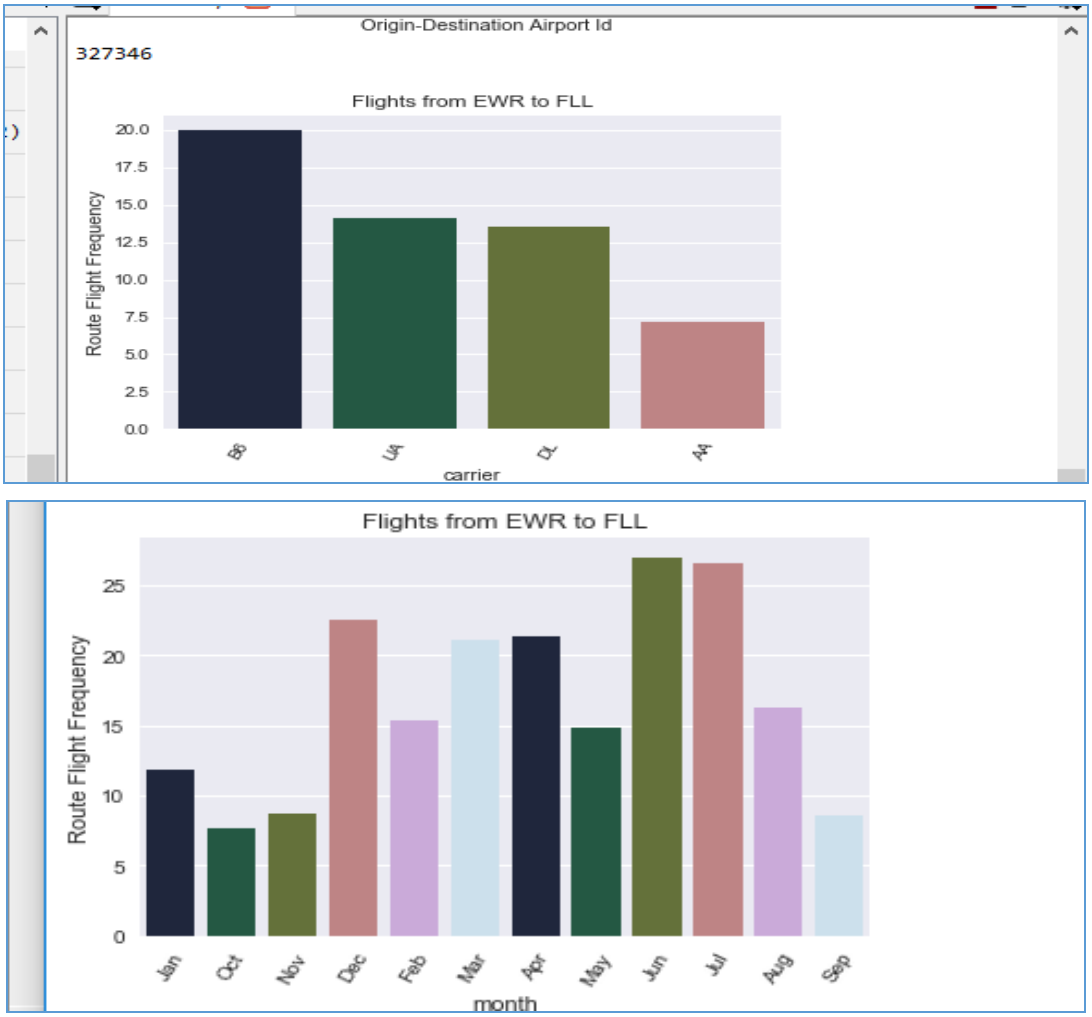


Figure 6: Flights From One Station to other Station and there Congestion

V. Conclusion and Future Scope

Most of the studies that have been proposed the last years and focus on the development of predictive models using supervised ML methods and Regression algorithms aiming to predict valid disease outcomes.

Received: 18 Mar 2019/Accepted: 17 Apr 2019

Based on the analysis of their results, it is evident that the integration of multidimensional heterogeneous data, combined with the application of different techniques for feature selection and regression can provide promising tools for inference in the cancer domain. The XGBoost is used in the analysis of this paper because XGBoost is one of the most popular machine learning algorithms these days. Regardless of the type of prediction task at hand; regression or classification. It has become the state-of-the-art machine learning algorithm to deal with structured data.

References

[1] https://en.wikipedia.org/wiki/Flight_cancellation_and_delay

[2] R. R. Clewlow, I. Simaiakis, and H. Balakrishnan, "Impact of arrivals on departure taxi operations at airports," 2010. CS229: AUTUMN, (2017).

[3] H. Balakrishnan, "Control and optimization algorithms for air transportation systems," Annual Reviews in Control, (2014).

[4] https://www.transtats.bts.gov/ONTIME/Departures.aspx.

Received: 18 Mar 2019/Accepted: 17 Apr 2019