Contents

**Predict – Flight Delay**

[Flight Delay 3](#_Toc527138588)

[Why is flight delay an important issue? 3](#_Toc527138589)

[Analysis of flight delay 4](#_Toc527138590)

[1. Cause of flight Delay 6](#_Toc527138591)

[6](#_Toc527138592)

[2. Process outline 8](#_Toc527138593)

[Business Problem 9](#_Toc527138594)

[Trend Analysis 9](#_Toc527138595)

[1. Air Time - Carriers 14](#_Toc527138596)

[2.Arrival and departure delays by carrier 15](#_Toc527138597)

[3. Number of trpis for each carrier 17](#_Toc527138598)

[4. City wise Delays 18](#_Toc527138599)

[5. State wise 19](#_Toc527138600)

[6. Relation between Number of planes scheduled and the delay caused by them. 20](#_Toc527138601)

[MODEL BUILDING 23](#_Toc527138602)

[1.Data preparation 23](#_Toc527138603)

[2.Model Development. 23](#_Toc527138604)

# Flight Delay

**A flight delay is when an airline flight takes off and/or lands later than its scheduled time. The Federal Aviation Administration (FAA) considers a flight to be delayed when it is 15 minutes later than its scheduled time.**

# Why is flight delay an important issue?

Flight delays have an impact on the following:

Airline Companies: It is an important issue for the airline as it is extra expenses for the company. The airline companies now-a -days don’t own the flights completely. The companies take them on a lease and would pay only the price for its time in air. So, if the plane was travelling for more time in air, this could be a very serious concern for the airlines.

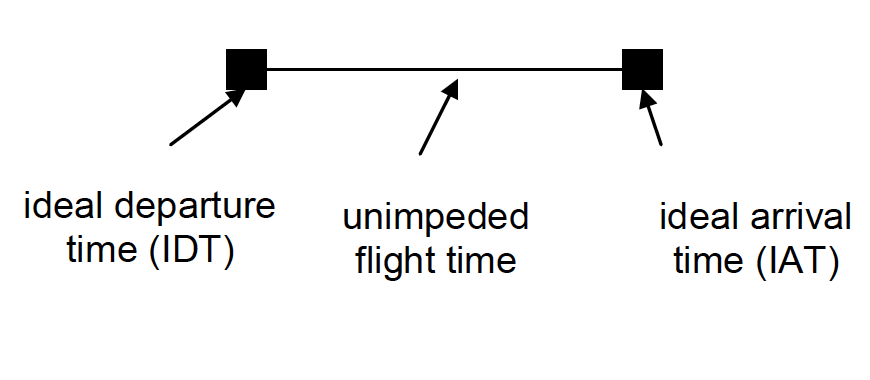
Passengers: We all know the time is money. Many people would feel unhappy with the carrier airline company if there is a delay in the flight schedule as it would lead to change in their own plans. If a person is late on a business trip, then the person is not only losing his personal time but also the employer’s time too. Many

Airport authorities: If everything happens as per the scheduled time, then it’s a happy situation to all, but if there is a delay in the schedule, then whole timings would be disturbed. As there is limited place on the runways and idle areas in for the airplanes, it’s a concern for the airport authorities.

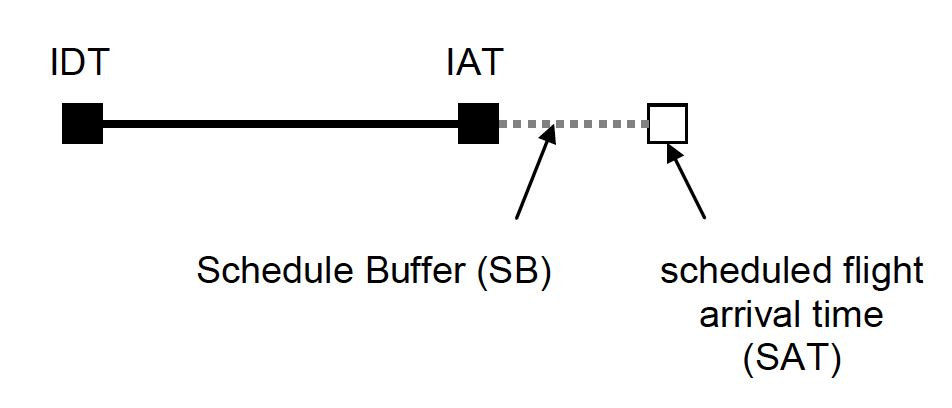
# Analysis of flight delay

Impact of flight delay

An airline starts the scheduling process of a flight by determining an ideal flight departure time or IDT. The IDT takes into account the preferred passenger travel times and also internal airline constraints such as those necessary to make proper crew efficient crew timing and fleet plans. As part of the process the airline chooses the most appropriate aircraft type from its fleet for the flight. Using the characteristics of that aircraft and assuming it flies at optimal conditions, unimpeded origin-to-destination trajectory, an ideal arrival time could be computed as shown in the Figure 1.

  
Figure 1: Ideal flight pattern

When we consider how congestion and delays can alter this situation, as illustrated in Figure2, the airlines typically increase scheduled flight times over unimpeded ones in order to consider the delays resulting from flight restrictions imposed to organise traffic, congestion and other such factors. This added time is referred to as schedule buffer or SB. Once an unimpeded flight time has been determined, the SB is computed from previous historical data.

  
Figure 2: Flight pattern when buffer is present

Congestion and delays affects airlines as well the passengers, but maybe in different ways. These have a definite impact on the airline costs. The passengers see increase in the time required for travel, experience discomfort and stress and may also sometimes face additional expenses for food and accommodation. The costs to airlines and passengers, some in the form of added expense and lost revenue and others in the form of decreased convenience and additional discomfort, these are the direct impacts of congestion and delay.

These direct congestion costs propagate through the rest of the economy, creating a third cost category. Any phenomenon that makes air transportation, more expensive leads to higher costs and lower efficiency in other segments, e.g. manufacturing, retail, etc. The added costs and reduced profits of any industry that depends on air travel, and the resulting impact on its customers, constitute the indirect impact of flight congestion and delay.

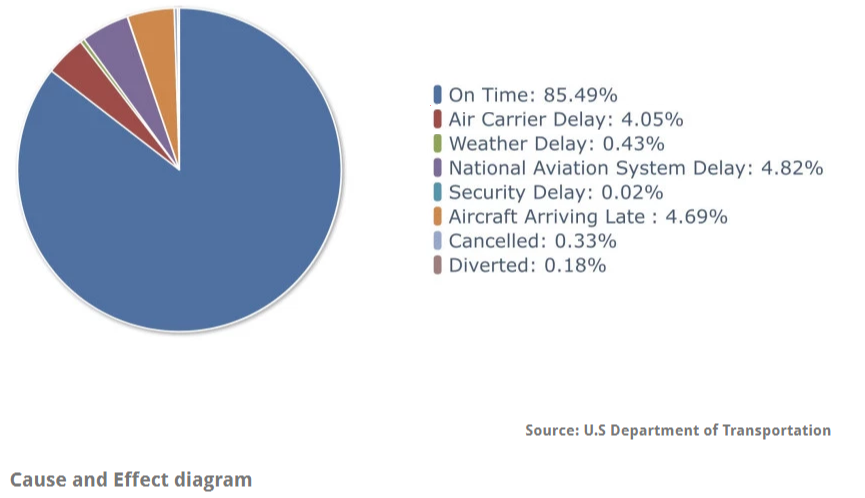
Flight delays affected passenger travel decisions and airline fares. It was found that flight delays by an airline on a route affected both passenger demand and the average fare on that route, hence the passenger demand reduced and average fares were increased. Experiments found that flight delays posed significant costs to both consumers and airlines, with producer costs being about three times the size (per passenger) of consumer costs. Calculations showed that the impact of reducing delays on welfare were very likely to be substantial. Although these positive welfare effects should be balanced against the capital expenditures necessary to reduce delays, there might be positive net benefits. Since the benefits for delay reductions were shared by passengers and airlines, the cost of paying for infrastructure improvements to reduce delays should have been likewise shared. Since airlines are the main beneficiaries from delay reductions, they should be assessed with the majority of the costs.

## Cause of flight Delay

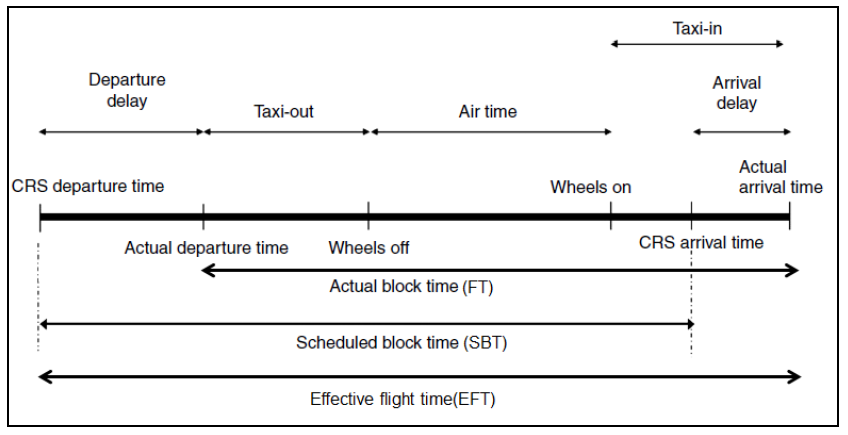
Here is a fishbone diagram representing the cause and effects of flights being delayed.

# Image result for fishbone for flight arrival delay

The Pie Chart below shows that there are only 85.49% of flights running as per the scheduled timing and there is also reason on why the rest couldn’t follow the same and was expressed in terms if percentages.



## Process outline



The picture here gives us a better understanding of various kinds of time metrics that would be calculated.

The most profitable model for an airline would be to fly when the plane is full and if the required passenger target is not met then the flight should be cancelled.

There are variety analyses a variety of cost components caused by flight delays, including cost to airlines, cost to passengers, cost of lost demand, as well as the indirect impact of delay on the US economy.

The direct costs imposed on the airline industry and its customers, flight delays have indirect effects on economies. Specifically, inefficiency in the air transportation sector increases the cost of doing business for other sectors, making the associated businesses less productive. The standard airline time measure gate to gate.

# Business Problem

We saw that the flight delays are inevitable and cannot be completely stopped. There are some scenarios where the delays are dependent on the nature, some others can be controlled by making some changes and a few things cannot be controlled. So we were clear that Flight delays cannot be eradicated completely but predicting the flight delays can help us to plan the schedule better. So the problem is to predict the flight delay.

# Trend Analysis

Now let us check it with the real time and start analysing the same with numbers. For that we need data. Data has been collected from United States Government website for the analysis.

The data was obtained from Bureau of Transportation Statistics and nearly 35 variables were removed since they didn’t not have any values present and it was impossible to impute. The target variable was not present directly. and it was a calculated, using Arrival Delay , Departure Delay. We have consider the data for one year, i.e. twelve months from June – 2017 to May – 2018 from publicly available website from United state Department of Transportation.

In the data that we collected we have the following variables.

**Variable Description**

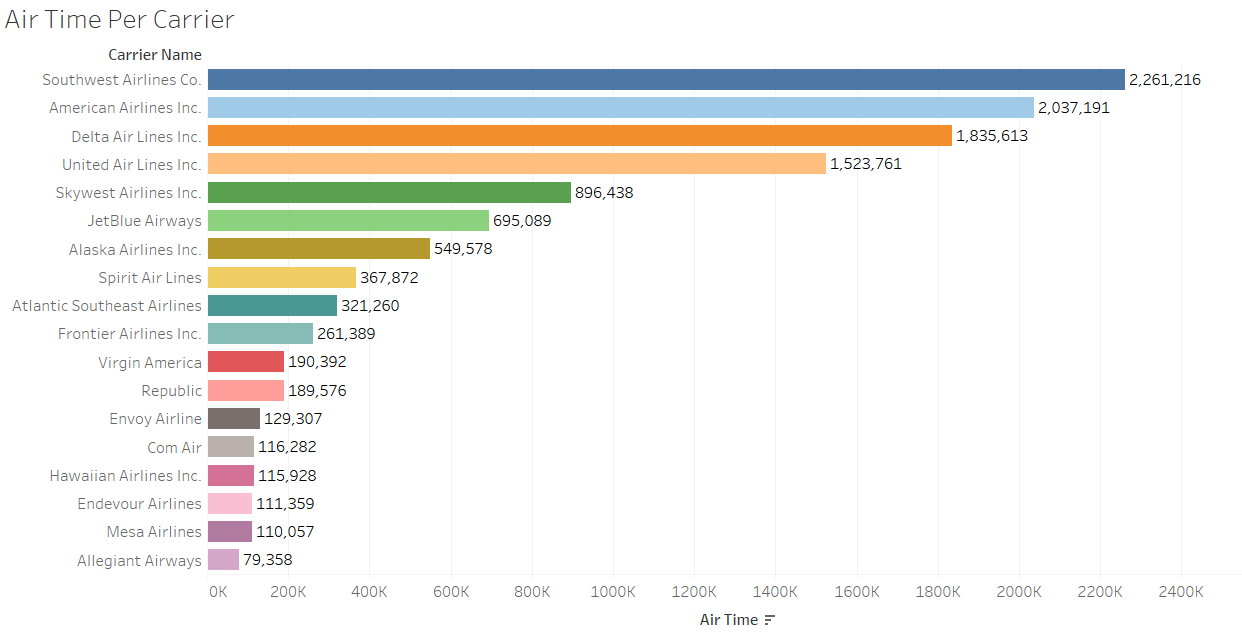
|  |  |
| --- | --- |
| **Name of Column** | **Description** |
| Year | Year |
| Quarter | Quarter (1-4) |
| Month | Month |
| DayofMonth | Day of Month |
| DayOfWeek | Day of Week |
| FlightDate | Flight Date (yyyymmdd) |
| UniqueCarrier | Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years. |
| AirlineID | An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation. |
| Carrier | Code assigned by IATA and commonly used to identify a carrier. As the same code may have been assigned to different carriers over time, the code is not always unique. For analysis, use the Unique Carrier Code. |
| TailNum | Tail Number |
| FlightNum | Flight Number |
| OriginAirportID | Origin Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused. |
| OriginAirportSeqID | Origin Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time. |
| OriginCityMarketID | Origin Airport, City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market. |
| Origin | Origin Airport |
| OriginCityName | Origin Airport, City Name |
| OriginState | Origin Airport, State Code |
| OriginStateFips | Origin Airport, State Code |
| OriginStateName | Origin Airport, State Name |
| OriginWac | Origin Airport, World Area Code |
| DestAirportID | Destination Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused. |
| DestAirportSeqID | Destination Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time. |
| DestCityMarketID | Destination Airport, City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market. |
| Dest | Destination Airport |
| DestCityName | Destination Airport, City Name |
| DestState | Destination Airport, State Code |
| DestStateFips | Destination Airport, State Fips |
| DestStateName | Destination Airport, State Name |
| DestWac | Destination Airport, World Area Code |
| CRSDepTime | CRS Departure Time (local time: hhmm) |
| DepTime | Actual Departure Time (local time: hhmm) |
| DepDelay | Difference in minutes between scheduled and actual departure time. Early departures show negative numbers. |
| DepDelayMinutes | Difference in minutes between scheduled and actual departure time. Early departures set to 0. |
| DepDel15 | Departure Delay Indicator, 15 Minutes or More (1=Yes) |
| DepartureDelayGroups | Departure Delay intervals, every (15 minutes from <-15 to >180) |
| DepTimeBlk | CRS Departure Time Block, Hourly Intervals |
| TaxiOut | Taxi Out Time, in Minutes |
| WheelsOff | Wheels Off Time (local time: hhmm) |
| WheelsOn | Wheels On Time (local time: hhmm) |
| TaxiIn | Taxi In Time, in Minutes |
| CRSArrTime | CRS Arrival Time (local time: hhmm) |
| ArrTime | Actual Arrival Time (local time: hhmm) |
| ArrDelay | Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers. |
| ArrDelayMinutes | Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0. |
| ArrDel15 | Arrival Delay Indicator, 15 Minutes or More (1=Yes) |
| ArrivalDelayGroups | Arrival Delay intervals, every (15-minutes from <-15 to >180) |
| ArrTimeBlk | CRS Arrival Time Block, Hourly Intervals |
| Cancelled | Cancelled Flight Indicator (1=Yes) |
| Diverted | Diverted Flight Indicator (1=Yes) |
| CRSElapsedTime | CRS Elapsed Time of Flight, in Minutes |
| ActualElapsedTime | Elapsed Time of Flight, in Minutes |
| AirTime | Flight Time, in Minutes |
| Flights | Number of Flights |
| Distance | Distance between airports (miles) |
| DistanceGroup | Distance Intervals, every 250 Miles, for Flight Segment |
| DivAirportLandings | Number of Diverted Airport Landings |
| Arrival Delay | Derived Column |
| Departure Delay | Derived Column |
| DELAY | Target column |

We collected the data. Upon doing some exploratory data analysis we understood the data. In the process we saw some tends in the data. After observing the trends, we concluded some insights that would help us to move further in the solving the issue.

## 1. Air Time - Carriers

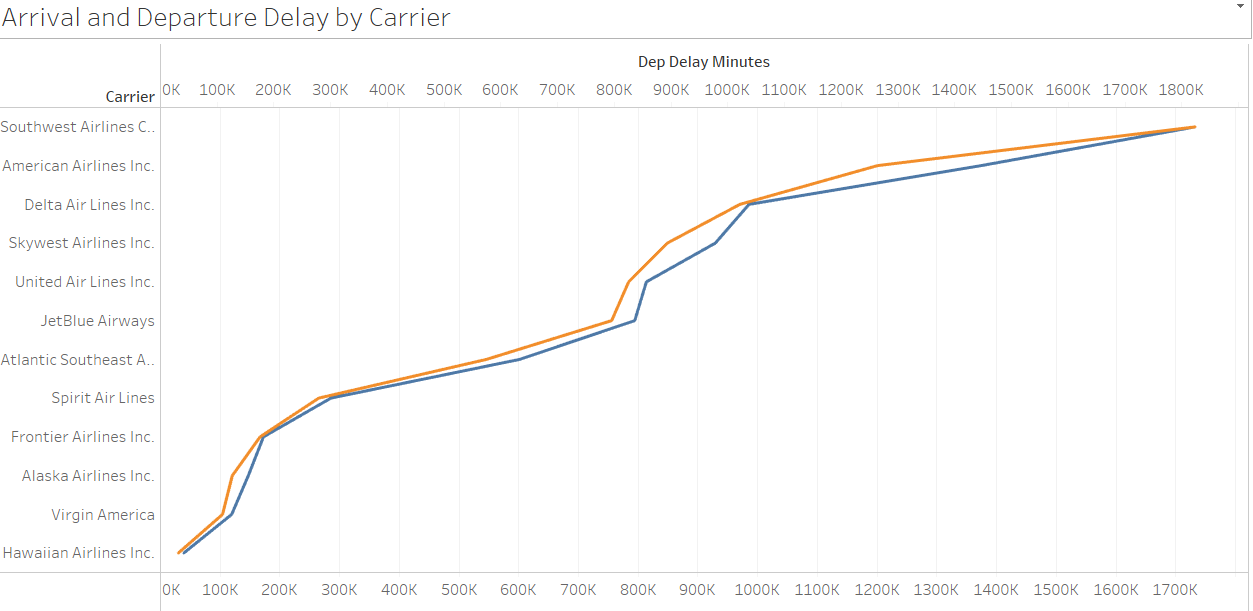
There are many airline companies that are operating in United States.

Now as per our data let us see the most popular one among them. For this we took the Air time as the parameter. In short Airtime is the time spent by the airplane in the air, the time starts from Wheels Off the runway of the origin airport to take-off and is counted till the airplane touches the destination runway for landing.



From the bar chart it is clearly evident that Southwest Airline Co. is leading the race confirming it to be most popular airline or carrier with 22,61,216 minutes. It list is then followed by American Airlines, Delta Airlines, United Airlines etc. The Allegiant Airlines has the least airtime than the rest of its competitors.

## 2.Arrival and departure delays by carrier



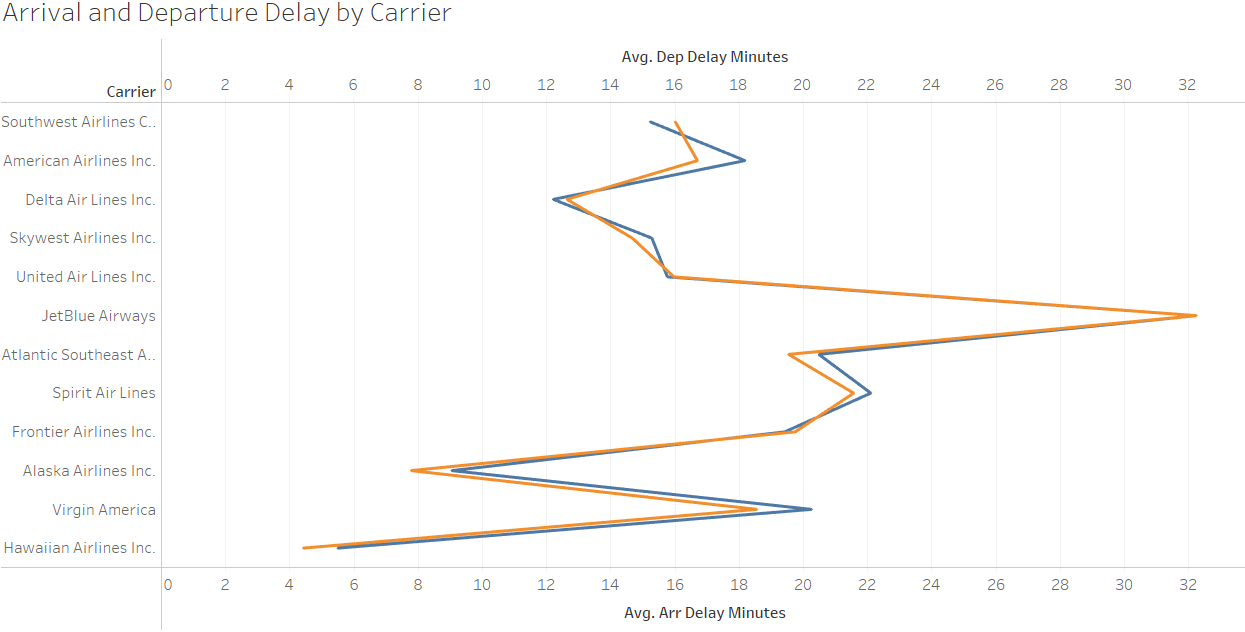
The sum of minutes of delay is taken as the parameter here to determine the flight delay timings.

In the chart we can see that Southwest Airline Co. topped the Arrival and Departure Delays.

It was already seen in the previous chart that it is the most popular airlines, but here it is observed that it is most carrier company with most delays too.

But how can the same company be preferred than its competitors?

The answer to that question is explained in the next chart.

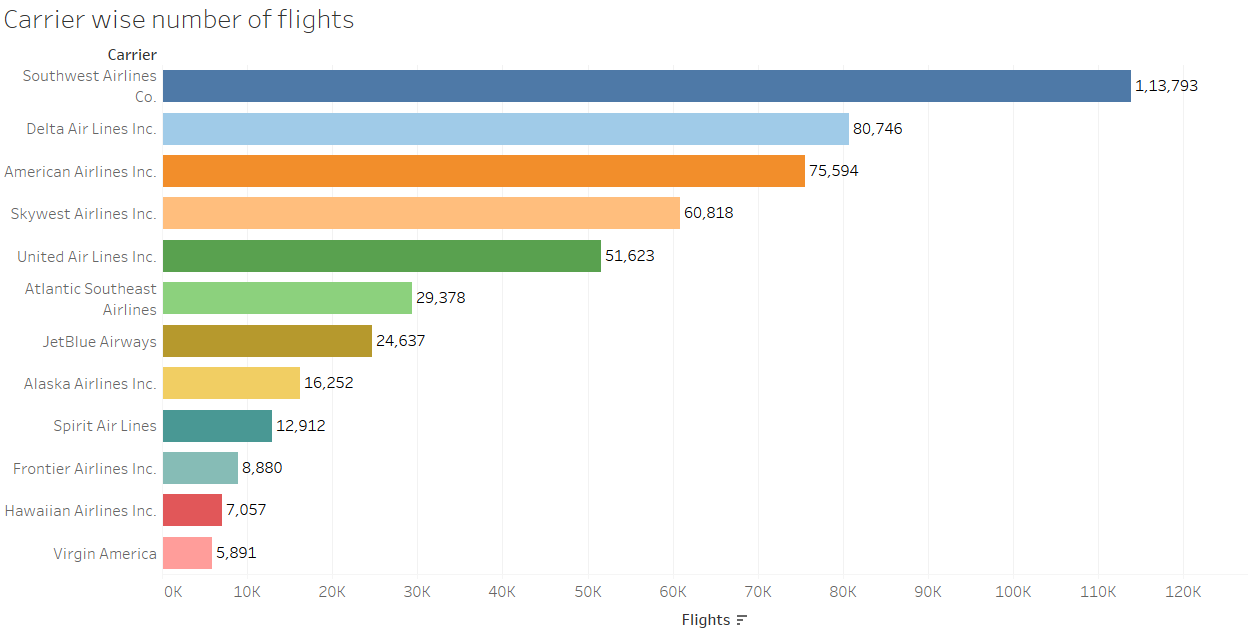


The average of minutes of delay is taken as the parameter here to determine the flight delay timings. Now we observe that Jet Blue Airlines is taking the maximum delay when compared with the rest others.

We can infer that Southwest Airlines is having the maximum number of flights. It can be observed when see that the same company is having maximum delay in terms of sum of minutes though being a mediocre in average minutes of delay.

Let’s check it.

## Number of trpis for each carrier

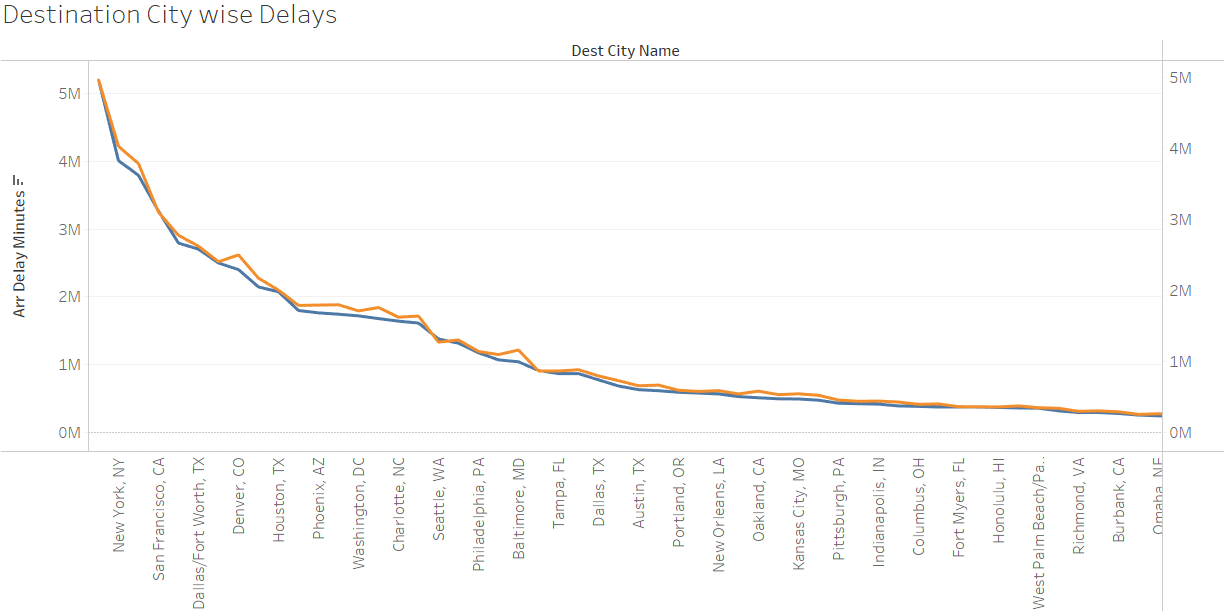


Yes, we can see in from the above chart that Southwest Airlines Co. has made the most number of flight journeys. The list is then followed by Delta Airline, American Airlines, SkyWest Airlines etc. Virgin Airlines has been observed with least number of trips.

Let us look into some more aspects.

The Air traffic changes in accordance with City.

## City wise Delays

When observed closely we can see that the following chart as:

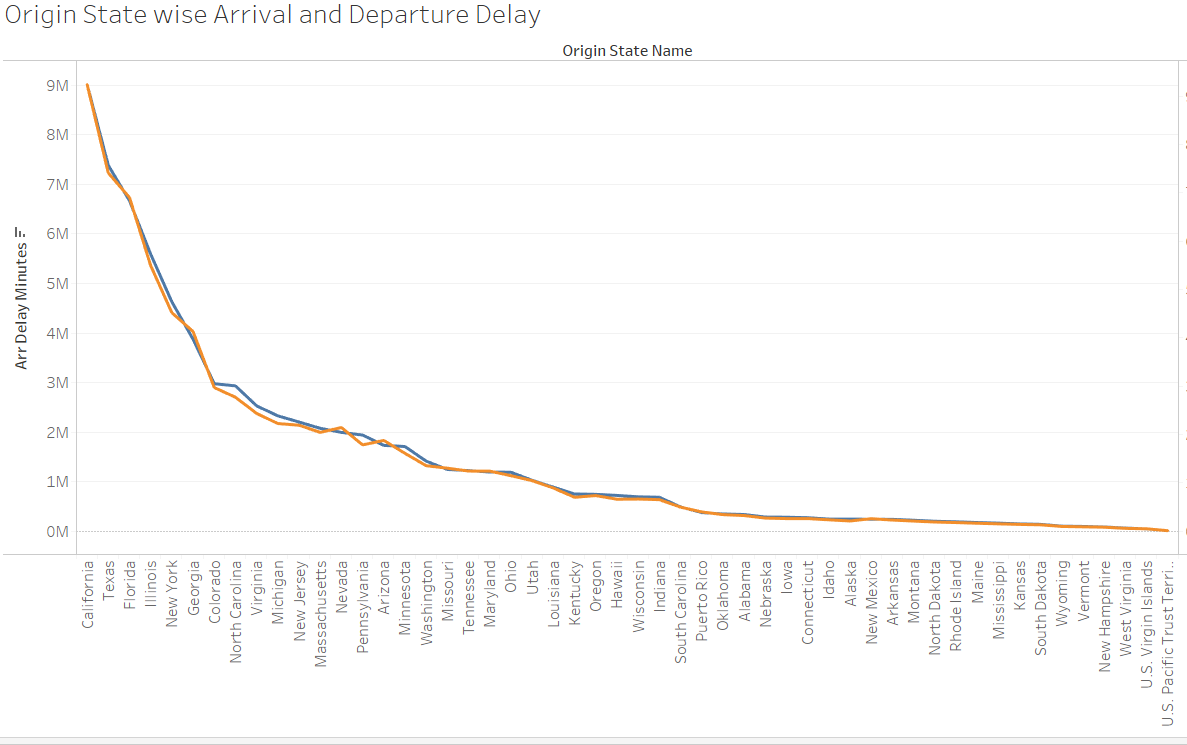
The New York city tops the list in terms of Arrival Delay and Destination Delay.

* The New York city airport has recorded 52,09,084 minutes of arrival delay i.e. 86,818 hours for all the flights collectively landed.
* The New York city airport has recorded 49,66,014 minutes of arrival delay i.e. 82,767 hours for all the flights collectively took off.

Its later followed by the cities of San Francisco, Dallas, Denver etc.

Now we observe the same at state level.

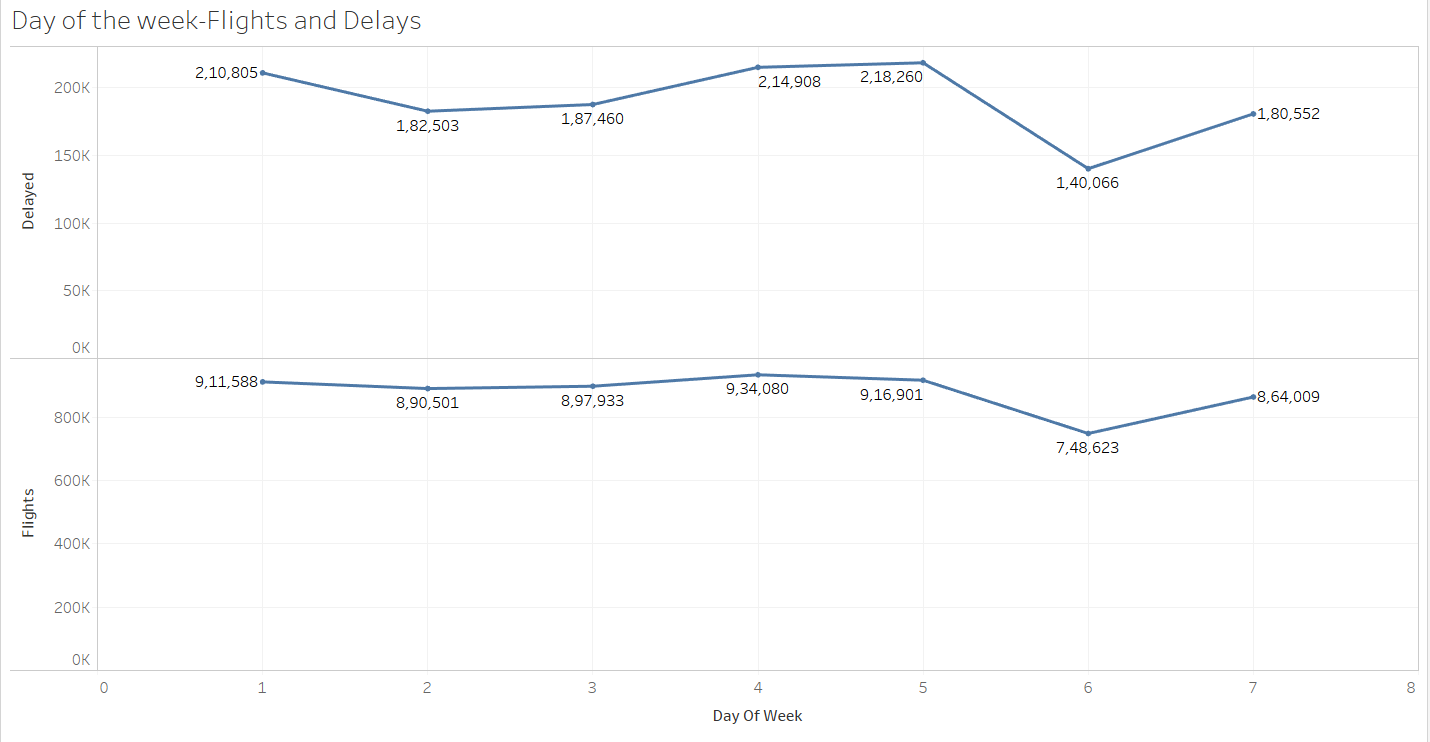
## State wise



The state of California tops the list in terms of Arrival Delay and Destination Delay.

* The California state has recorded 90,20,749 minutes of arrival delay i.e. 1,50,345 hours for all the flights collectively landed.
* The California state has recorded 92,42,632 minutes of arrival delay i.e. 1,54,043 hours for all the flights collectively took off. Its later followed by the states of Texas, Florida, Illinois etc.

## Relation between Number of planes scheduled and the delay caused by them.

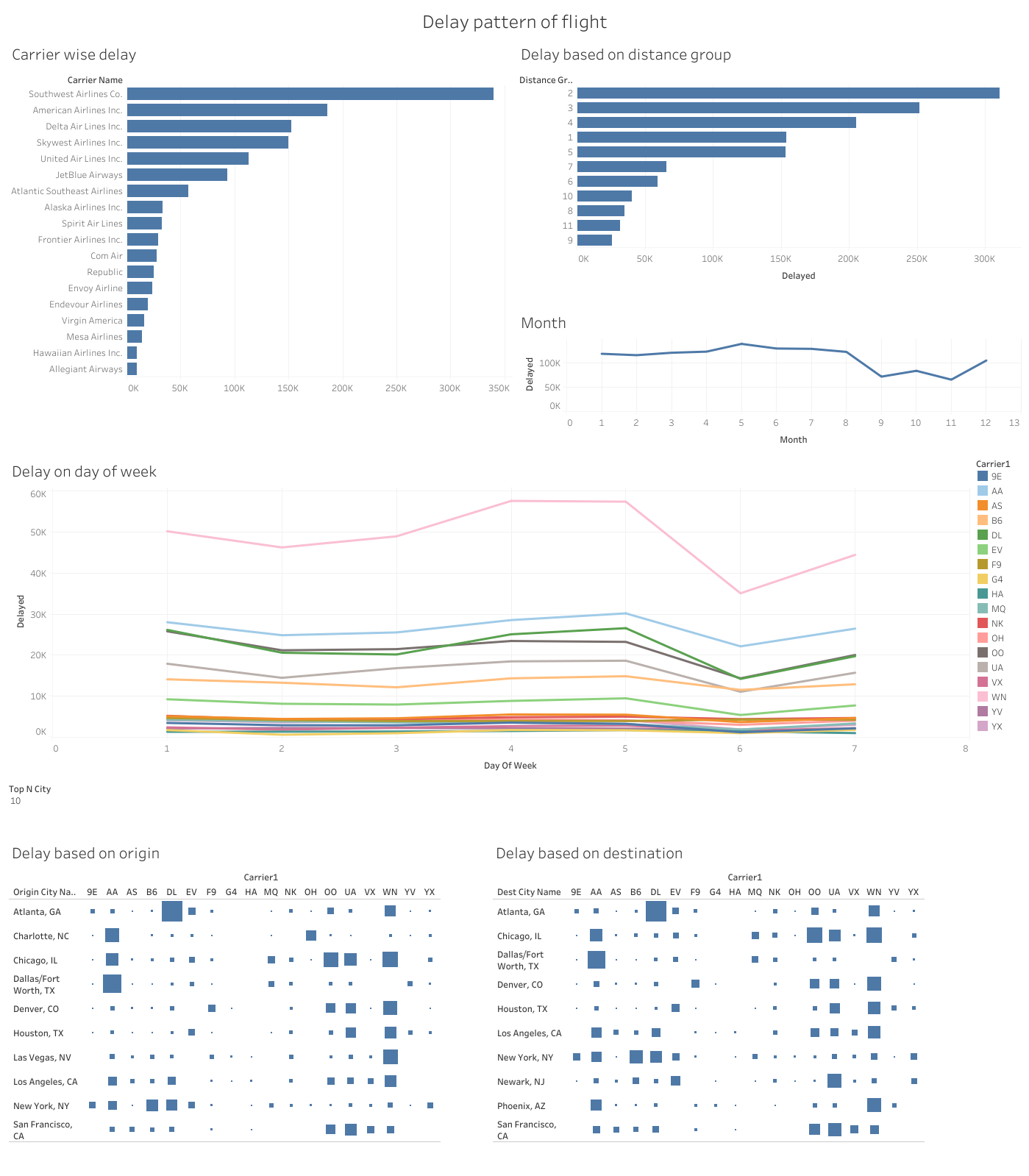


We could see that the number of planes scheduled and the number of planes getting delayed are almost following the same pattern. The sharp decline in delays when there are less number of flights scheduled on Day6 shows then clear picture of the observation. The decline and immediate raise strengthens the same. On the contrary we could also observe that there is not much decline between Day1 and Day2 in the number of flights when compared with the flights being delayed.

As the dataset was too huge we just had all the observations so far observations only on a part of data.

Now we applied it for the whole data and found that all our insights still continued to be in the very same trend and there were a few discussions that have been shared below.

.



* In the carrier wise delay Southwest Airlines Co. having the highest number of delay carrier wise and followed by American Airline Inc.
* Based on the distance group by delay group 2 having the highest delayed group.
* In May month highest delayed group is 2 ( 250-500 Miles ) followed by group 3 (500 – 750 Miles).
* On the Friday for each airline there is highest delay is on Friday. Within that Southwest Airlines Co. having the highest delay.
* Delayed based on the Origin City
* Highest delayed based on the Atlanta origin city is the Delta Air Line Inc. carrier. Followed by Dallas origin city is the American Airline Inc.
* Delayed based on the Destination City.
* Highest delayed based on the Atlanta destination city is the Delta Air Line Inc. carrier. Followed by Dallas destination city is the American Airline Inc.
* If we see the pattern origin city, destination city and airlines are repeated in both delayed based on Origin city and destination city.

# MODEL BUILDING

With all our insights we started preparing the data for developing our model.

## 1.Data preparation

[Link](https://transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time)

The data was obtained from Bureau of Transportation Statistics and nearly 35 variables were removed since they didn’t not have any values present and it was impossible to impute. The target variable was not present directly. and it was a calculated, using Arrival Delay , Departure Delay. We have consider the data for 1 year from June – 2017 to May – 2018 from publicly available website from United state Department of Transportation.0

## 2.Model Development.

Based on the business problem and attributes contributing the flight delay and the data which can be provided by customer to predict the flight will delay or not. We have removed the variables as we found them duplicating or irrelevant. The variables that have been dropped are listed below.

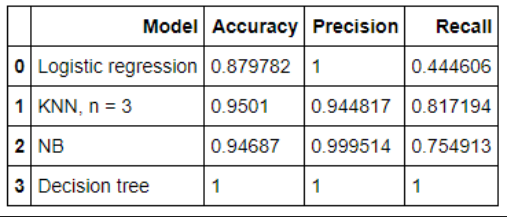
|  |  |
| --- | --- |
| FlightDate | Flight Date (yyyymmdd) |
| ActualElapsedTime | Elapsed Time of Flight, in Minutes |
| ArrDel15 | Arrival Delay Indicator, 15 Minutes or More (1=Yes) |
| ArrDelay | Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers. |
| ArrTime | Actual Arrival Time (local time: hhmm) |
| ArrTimeBlk | CRS Arrival Time Block, Hourly Intervals |
| CRSArrTime | CRS Arrival Time (local time: hhmm) |
| CRSDepTime | CRS Departure Time (local time: hhmm) |
| CRSElapsedTime | CRS Elapsed Time of Flight, in Minutes |
| DepDel15 | Departure Delay Indicator, 15 Minutes or More (1=Yes) |
| DepDelay | Difference in minutes between scheduled and actual departure time. Early departures show negative numbers. |
| DepDelayMinutes | Difference in minutes between scheduled and actual departure time. Early departures set to 0. |
| DepDelayMinutes | Difference in minutes between scheduled and actual departure time. Early departures set to 0. |
| DepTime | Actual Departure Time (local time: hhmm) |
| DepTimeBlk | CRS Departure Time Block, Hourly Intervals |
| Diverted | Diverted Flight Indicator (1=Yes) |
| TailNum | Tail Number |
| TaxiIn | Taxi In Time, in Minutes |
| TaxiOut | Taxi Out Time, in Minutes |
| UniqueCarrier | Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years. |
| Year | Year |

As the dataset was huge ( about 66 lakhs rows and 50 columns )

The dataset was huge that is why we have taken the test size is 0.60 ( 60% ) of the whole dataset.

Four model i.e., Logistic Regression, K-Nearest Neighbour, Naïve Bayes and Decision Tree have been opted to run on the training dataset.

After running the Models on the training dataset. We have received test accuracy given below.



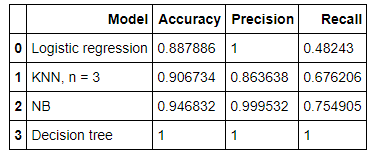
It was a fine fit with accuracy ranging from nearly 88% to 100%.

* The accuracy of Logistic Regression is around 88%
* The accuracy of K-Nearest Neighbour is around 95%
* The accuracy of Naïve Bayes is around 94.6%
* The accuracy of Decision Tree is around 100%, which is like too good to be true.

All the models gave very promising accuracy stating that the models were very well fit on the training datasets. Though this is good sign we cannot confirm the results unless tested upon the test dataset.

Now, let’s check out the model performance on the testing dataset.

After running the Models on the dataset. We have received test accuracy given below.



* The accuracy of Logistic Regression is around 88%
* The accuracy of K-Nearest Neighbour is around 90%
* The accuracy of Naïve Bayes is around 94.6%
* The accuracy of Decision Tree is around 100%, which is like too good to be true.

All the models performed well on the testing data too. When compared we can observe that

* Logistic Regression accuracy was almost the same on the training and testing data
* K-Nearest Neighbour fumbled the most giving a difference on 5% accuracy but yes we cannot tell that it over fit the test data.
* On the other hand, Naïve Bayes also performed as good as it performed on the training data
* Last but never the least, Decision Tree has predicted the most accurate result the one can ever get. It gave 100% accuracy on the training and testing data as well.