International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

Multinomial Logistic Regression Model for Predicting Flight Arrival & Delay

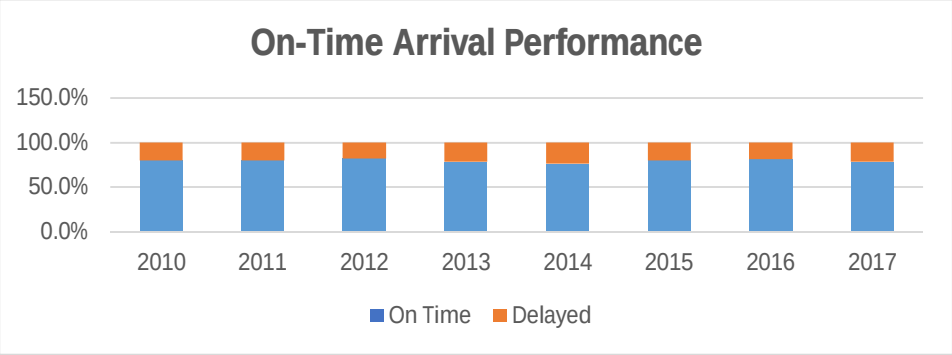
Radhaiah Konidina[1], M. Venkataramanaiah[2]
[1, 2] *Department of Statistics, S.V. University, Tirupati.*

Abstract: ***Transportation has transformed the human evaluation by adding another dimension: Speed to life and now in the modern generation, every delayed second counted as a missed opportunity. All forms of transportation operators showing momentous efforts to keep up to the schedules and make their customers happy. Aviation is one of such industry, in this article we tried to find the major reasons behind Aircraft arrival delays and predict the average arrival delay by building a multinomial logistic regression.***

Keywords: Arrival delay, Logistic Regression, Multinomial Logistic Regression, Ordinal Response and Nominal Response.

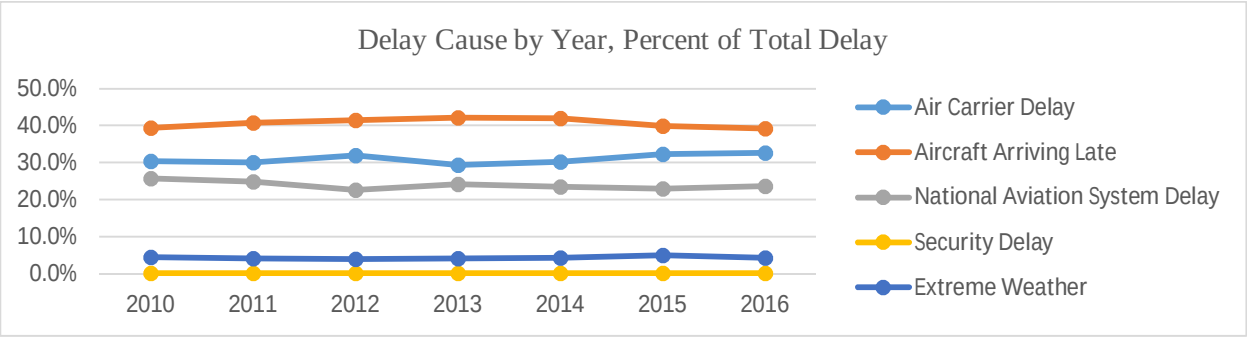
I. INTRODUCTION

The aviation industry in has emerged as one of the fastest growing industries during the last few years. The Aviation industry has piloted in a new era of expansion, driven by factors such as low-cost carriers (LCCs), modern airports, Foreign Direct Investment (FDI) in domestic airlines, advanced information technology (IT) interventions and growing emphasis on regional connectivity. Aviation industry may help the economy of the country as they provide employment to a larger number of people with heavy pay packages. The profitability of the aviation is all centered on schedules; airlines operate flights on fixed schedules regardless of flight loads. They may not make for a good business with means of empty planes as they mean no profits. Airlines adjust their schedules to capture the most profitable market but air travel is seasonal and volatile. The most profitable model for an airline would be to fly when the plane is fully occupied and if the required passenger target is not met then the flight should be canceled. When you buy a ticket with an airline you expect that flight to depart at the scheduled time and the scheduled day you booked a ticket for. For this reason, an airline must adhere to their advertised route and schedule. Scientific data about the cost of flight delays is still limited but we all experience the effects when a flight we are about to depart is delayed. In 2007 the FAA-sponsored some US universities to study flight delay impacts in the United States. This report analyses a variety of cost components caused by flight delays, including the cost to airlines, cost to passengers, cost of lost demand, as well as the indirect impact of delay on the US economy. The project team estimates that the total cost of all US air transportation delays in 2007 was $32.9 billion. The $8.3 billion airline component consists of increased expenses of the crew, fuel, and maintenance, among others. The $16.7 billion passenger component is based on the passenger time loss due to schedule buffer, delayed flights, flight cancellations, and missed connections. The $3.9 billion cost from lost demand is an estimate of the welfare loss incurred by passengers who avoid air travel as the result of delays. In addition to these 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. In the below table we can observe the percentage of on-time vs delayed flights in the US from the year 2010 to 2017, in almost all the years around 20% of the flights are delayed.



SOURCE: Bureau of Transportation Statistics, Airline Service Quality Performance 234

There are four possibilities for incoming flights namely on-time arrival, delayed, Cancelled and diverted. According to the US Bureau of transportation, the largest most of flight delays were caused by late arrival of incoming aircraft. If an aircraft is late at one destination very rarely does the next service for that aircraft leave on time, the delay is reactionary. But what causes these delays in the first place? In the below table we can observe the distribution of causes for the delays in the US from 2010 to 2016, from the chat we can see that Air Carrier Delay is the most frequent cause for a plane delay. Aircraft Maintenance, late crew, toilet cleaning and catering etc., are few of the things we can consider as Air Carrier delays. Next to Air Carrier Delay we have connecting Aircraft Arriving late as other reason for the delay.



In this paper, we tried to find the factors that may affect the flight arrival status and fit a model that can predict the flight arrival status. We segmented all the arrival status of the flights into different groups and used a Multinomial Logistic Regression model to predict the outcome.

II. DATA AVAILABILITY

BTS data collections include traffic, passenger flow, employment, financial condition, and on-time performance of commercial aviation; the Commodity Flow Survey; trans-border movement of freight by mode of transportation; a census of ferry operations, precursor safety data for transit operations, and data on near misses and equipment failures in offshore operations. We collected 'on-time performance of commercial aviation' for the period of Jan-2017 to Aug-2017 from BTS for the analysis. In this data, we have information like Flight Number, Origin City, Destination City, Departure Time, Departure Delay, Arrival time, Arrival Delay, Total Air Time, Total Distance and Total Flights on the move in the same time etc.,

III. METHODOLOGY

The dependent variable in most regression models is numerical, measured usually on a ratio scale. But in many applications, the dependent variables are nominal in the sense that they denote categories, such as male or female. How do we model such nominal variables? Can we use the traditional regression techniques or do we need specialized techniques? Regression models involving nominal scale variables are an example of a broader class of models known as qualitative response regression models. There are a variety of such models is the binary or dichotomous or dummy dependent variable regression models. In these scenarios, our primary objective is to estimate the probability of the values of the explanatory variables. In developing such a probability function, we need to keep in mind two requirements:

As *Xi*, the value of the explanatory variable(s) changes, the estimated probability always lies in the 0–1 interval, and

The relationship between *Pi* and *Xi* is nonlinear, that is, "one which approaches zero at slower and slower rates(slower?? Rate??) as *Xi* gets small and approaches one at slower and slower rates as *Xi* gets very large".

The logistic Regression models satisfy these requirements.

A. Logistic Regression Model

This is the most important model for categorical response data. It is used increasingly in a wide variety of applications. Early uses were in biomedical studies but the past 20 years have also seen much use in social science research and marketing. Recently, logistic regression has become a popular tool in business applications. Some *credit-scoring* applications use logistic regression to model the probability that a subject is creditworthy. For instance, the probability that a subject pays a credit card bill on time may use predictors such as the size of the bill, annual income, occupation, mortgage and debt obligations, the percentage of bills paid on time in the past, and other aspects of an applicant's credit history. A company that relies on catalog sales may determine whether to send a catalog to a potential customer by modelling the probability of a sale as a function of indices of past buying behavior. Another

area of increasing application is genetics. For instance, one recent article (J. M. Henshall and M. E. Goddard, Genetics 151:885_894,1999) used logistic regression to estimate quantitative trait loci effects, modelling the probability that an offspring inherits an allele of one type instead of another type as a function of phenotypic values on various traits for that offspring. Another recent article (D. F. Levinson et al., Amer. J. Hum. Genet., 67:652_663, 2000) used logistic regression for analysis of the genotype data of affected sibling pairs (ASPs) and their parents from several research centers. The model studied the probability that ASPs have identity by-descent allele sharing and tested its heterogeneity among the centers.

B. Parameters in Logistic Regression

For a binary response variable *Y* and an explanatory variable *X*, Let $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$. the logistic regression model is

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

Equivalently, the log odds, called the logit, has a linear relationship

$$\text{Logit}[\pi(x)] = \log(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta x$$

This equates the logit link function to the linear predictor.

C. Interpreting β : Odds, Probabilities, and Linear Approximations

The sign of β determines whether $\pi(x)$ is increasing or decreasing as *x* increases. The rate of climb or descent increase as $|\beta|$ $increases$; $as\ \beta \to 0$ the curve flattens to a horizontal straight Line. When$\beta = 0$, *Y* is independent of *X*. For quantitative *x* with $\beta > 0$, the curve of $\pi(x)$ has the shape of the cdf of the logistic distribution. Since the logistic density is symmetric, $\pi(x)$ approaches 1 as the same rate that approaches 0. Exponentiation both sides of $\pi(x)$ shows that the odds are an exponential function of *x*. This provides a basic interpretation for the magnitude of β: the odds increase multiplicatively by e^{β} for every 1-unit increase in *x*. In other words e^{β} is an odds ratio, the odds at $X = x + 1$ divided by the odds at $X = x$.

D. Inference for Logistic Regression

By Wald's (1943) asymptotic results for ML estimators, parameter estimators in logistic regression models have large-sample normal distributions. Thus, an inference can use the (Wald, likelihood-ratio, score) triad of methods.

For the model with a single predictor, Logit $[\pi(x)] = \alpha + \beta x$,

Significance tests focus on *H*0: $\beta = 0$, the hypothesis of independence. The Wald test uses the log likelihood at β, with test statistic $z = \hat{\beta}/\text{SE}$ or its square; under *H*0 , z^2 is asymptotically $\varkappa^2 1$. The likelihood-ratio test uses twice the difference between the maximized log likelihood atβ, and at β =0 and also has an asymptotic $\varkappa^2 1$ null distribution. The score test uses the log likelihood at β =0 through the derivative of the log likelihood (i.e., the score function) at that point. The test statistic compares the sufficient statistic for β to its null expected value, suitably standardized [*N*(0, 1) or $\varkappa^2 1$.] .

For large samples, the three tests usually give similar results. The likelihood-ratio test is preferred over the Wald. It uses more information since it incorporates the log likelihood at *H*0 as well as atβ. When $|\beta|$ is relatively large, the Wald test is not as powerful as the likelihood-ratio test and can even show aberrant behavior [Hauck and Donner (1977) and Problem 5.38].

Till now we discussed building a logistic regression model for binary response variables with binomial GLMs. Multi-category responses use multinomial GLMs, where we can generalize logistic regression for multinomial. In the multinomial examples, we may have two possibilities nominal and ordinal.

E. Nominal Responses

Let *Y* be a categorical response with *J* categories. Multicategory (also called Polytomous) logit models for nominal response variables simultaneously describe log odds for all $\binom{J}{2}$ pairs of categories. Given a certain choice of *J* - 1 of these, the rest are redundant.

F. Baseline-Category Logits

Let $\pi_j(x) = P(Y = j|x)$ at a fixed setting **x** for explanatory variables, with $\sum \pi_j(x) = 1$. For observations at that setting, we treat the counts at the *J* categories of *Y* as multinomial with probabilities $\{\pi_1(x), \pi_2(x) \ldots\ldots \pi_j(x)\}$

Logit models pair each response category with a baseline category, often the last one or the most common one. The model

$$\text{Log}\frac{\pi_j(X)}{\pi_J(X)} = \alpha_j + \beta_j'(x) \quad \text{J=1, 2, 3, ----, j-1}$$

Simultaneously describes the effects of **x** on these J - 1 logits. The effects vary according to the response paired with the baseline. These J - 1 equations determine parameters for logits with other pairs of response categories, since

$$\text{Log}\frac{\pi_a(X)}{\pi_b(X)} = \text{Log}\frac{\pi_a(X)}{\pi_j(X)} - \text{Log}\frac{\pi_b(X)}{\pi_j(X)}$$

With categorical predictors, X^2 and G^2 goodness-of-fit statistics provide a model check when data are not sparse. When an explanatory variable is continuous or the data are sparse, such statistics are still valid for comparing nested models differing by relatively few terms (Haberman 1974a, pp.
372_373; 1977a)

G. Ordinal Responses: Cumulative Link Models

Cumulative logit models use the logit link. As in univariate GLMs, other link functions are possible. Let G^{-1} denote a link function that is the inverse of the continuous CDF G.

The *cumulative link* model

$$G^{-1}[P(Y \leq j|\ \mathbf{x}] = \alpha_j + \beta_j'(x)$$

links the cumulative probabilities to the linear predictor. The logit link function
$G^{-1}(u) = \log(u/1-u)$ is the inverse of the standard logistic CDF. As in the proportional odds model, effects of **x** in the model are assumed the same for each cutpoint, j=1, 2, 3, . . . J -1. we can show that this assumption holds when a linear regression for a latent variable Y * has standardized CDF G.

This Model results from discrete measurement of Y * from a location-parameter family having CDF $G(y^* - \beta'(x))$. The parameters $\{\alpha_j\}$ are category cut points on a standardized version of the latent scale. In this sense, cumulative link models are regression models, using a linear predictor $\beta'(x)$ to describe effects of explanatory variables on crude the ordinal measurement of Y *. Using -β rather than +β in the linear predictor merely results in change of sign of $\dot{\beta}$

IV. DATA COLLECTION

A. Flight On-time Performance

We have collected flight on-time performance data from The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS), they publish flight performance data within 30 days after the month end to the public. BTS collects data from various sources including their own data collection, data collection includes traffic, number of passengers, financial status and on-time performance of the flights. Below are few of the columns from the extracted data.

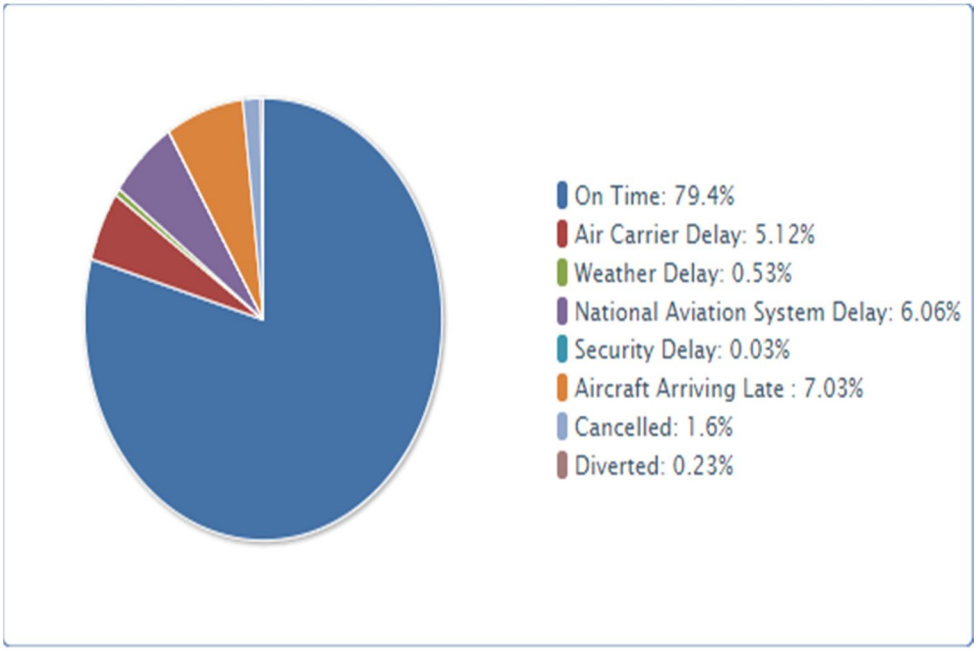
FlightNum	Flight Number
Origin	Origin Airport
OriginCityName	Origin Airport, City Name
OriginState	Origin Airport, State Code
OriginStateFips	Origin Airport, State Fips
OriginStateName	Origin Airport, State Name
OriginWac	Origin Airport, World Area Code
Dest	Destination Airport
DestCityName	Destination 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.

B. Weather Data

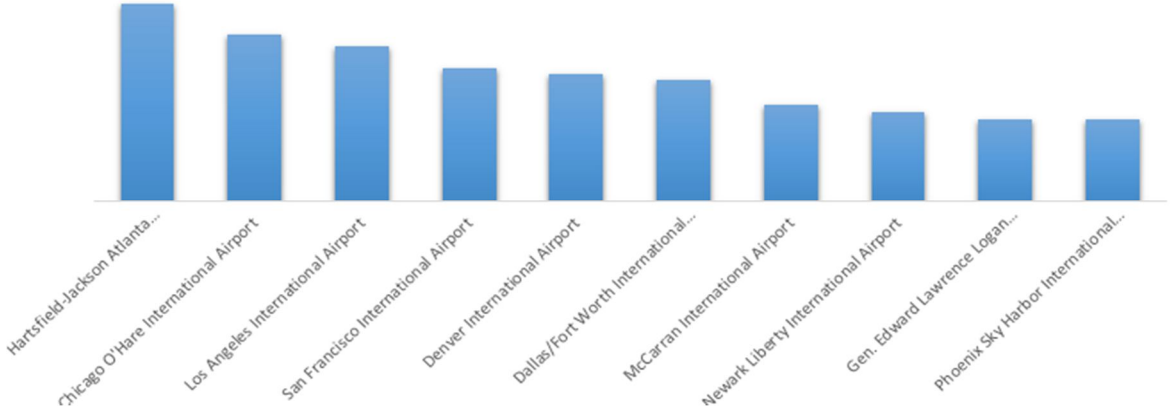
Historically we can observe that around 5% of the delayed flights are due to extreme weather, so we considered weather information in the prediction. Weather data has been collected from National Climatic Data Centre (NCDC) which is part of National Oceanic and Atmospheric Administration (NOAA). NCDC is responsible for historical weather monitoring, assessing, preserving and providing public access. NCDC provides access to sub-hourly (5-minute) data from the U.S. Climate Reference Network / U.S. Regional Climate Reference Network (USCRN/USRCRN) via anonymous ftp at:
ftp://ftp.ncdc.noaa.gov/pub/data/uscrn/products/subhourly01

V. EXPLANATORY DATA ANALYSIS

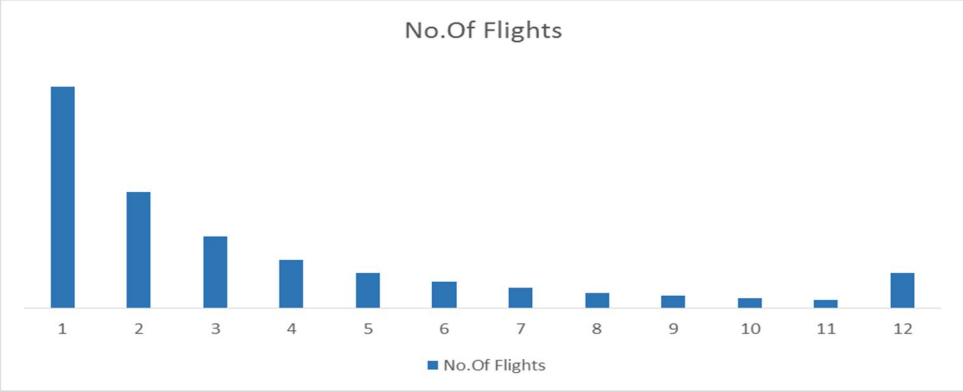
We conducted initial data analysis from the extracted data, from the below graph we can see that of all the 2017 US flights 79.4% of the flights arrive on time and around 20% of the flights are delayed, and if we check the reasons behind the flight delays Aircraft Arriving late is the major cause for a flight delay with 7.03% followed by National Aviation System Delay with 6.06% and Air Carrier Delay with 5.122%.



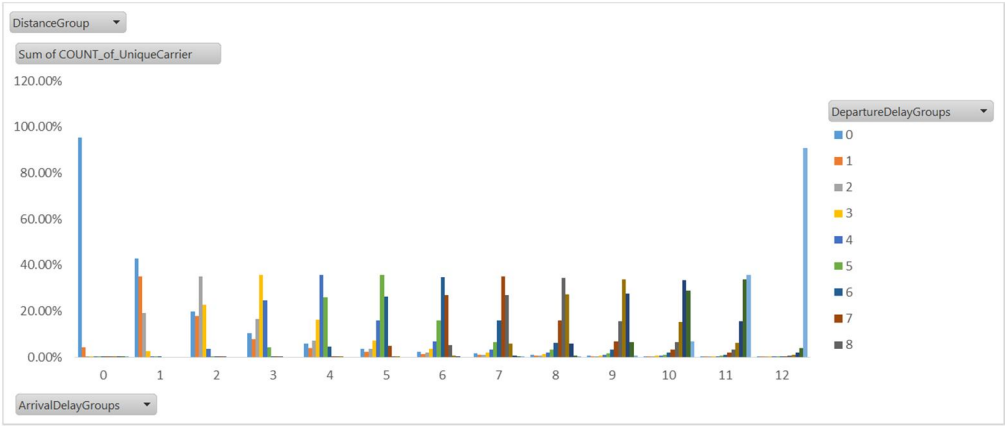
In the below graph we can see the number of delayed flights by Airport, and Hartsfield-Jackson Atlanta International Airport is in top place followed by Chicago O'Hare International Airport, Los Angeles International Airport and San Francisco International Airport. We know that Hartsfield-Jackson Atlanta International Airport is one of the busiest airports in the world, this will prone us to include the variable total number of flights arriving and departing from the airport to consider the traffic of the airport.



In the below chart we can see the total number of flights by the delayed group, most of the flights are in 15 minutes to 30 minutes delay group.

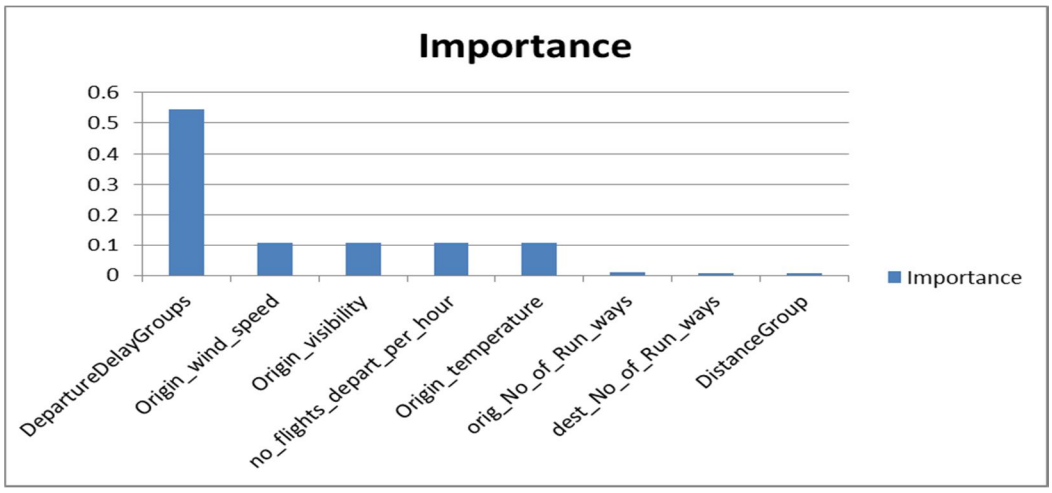


In the below chart we can observe the arrival delay with respect to the departure delay, about 95% of the aircrafts which departs on-time are arrived on-time, around 42% of the flights which departs 15 Minutes late are arriving on-time and around 0.06% of the flights which departs very late arrive on-time; this will give us evidence that departure delay is one of the main reason for the aircraft arrival delay and we can include this in our model. Around 91% of the flights which are in most delayed group arrive in the same delay group when the distance between the origin and destination is less, this delay percentage came down to 83% if the distance between the origin and destination is high, this will inspire us to include the distance between the origin and destination in the model.



VI. ORDINAL LOGISTIC REGRESSION

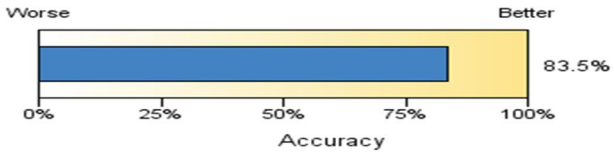
Here our dependent variable “Arrival Delay group” is an ordinal variable, so we can fit an ordinal logistic regression model.In the below table we can see the important variables considered in the model.



It can be seen the above chart that the variable Departure Delay group has high importance in the model and other variable have the same importance in the model. In the below image we can see the details of the model building process. We have around 83.5% accuracy in the prediction.

Target: ArrivalDelayGroups

Target	ArrivalDelayGroups
Measurement Level	Nominal
Probability Distribution	Multinomial
Link Function	Generalized logit



In the below heat map we can observe that the for Arrival Delay Groups 'On-time arrival' and the highest arrival delay group 12 we have around 99.9% and 91.1% accuracy respectively, for the other all groups there are around 35% accuracy, if we observe the classifications the delays are classified into the nearest delay group.Here may be by clubbing few delayed groups into one can help in improving the model accuracy.

Classification

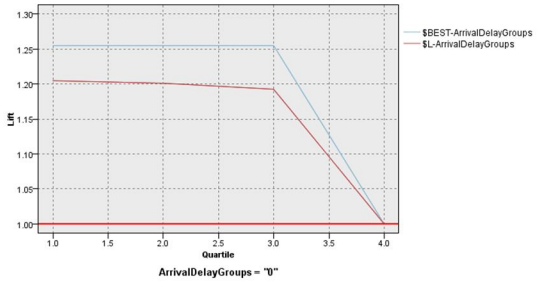
Target:ArrivalDelayGroups

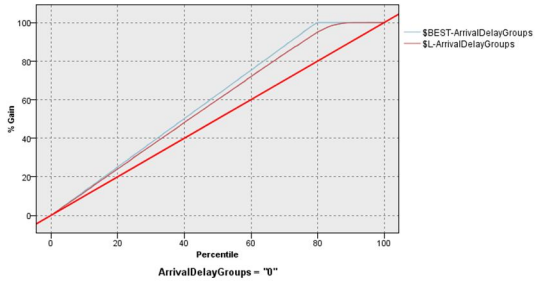
Overall Percent Correct =83.5%

Observed	Predicted												
	0	1	10	11	2	3	4	5	6	7	8	9	12
0	99.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
1	97.1%	0.0%	0.0%	0.0%	0.0%	2.7%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
10	0.8%	0.0%	31.9%	28.7%	0.0%	0.6%	1.1%	1.4%	2.2%	3.0%	7.6%	15.9%	6.9%
11	0.4%	0.0%	14.2%	35.0%	0.0%	0.3%	0.3%	0.9%	0.6%	2.6%	2.2%	6.2%	37.2%
2	72.5%	0.0%	0.0%	0.0%	0.0%	23.4%	3.6%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%
3	35.3%	0.0%	0.0%	0.0%	0.0%	35.8%	24.4%	4.1%	0.4%	0.0%	0.0%	0.0%	0.0%
4	17.3%	0.0%	0.0%	0.0%	0.0%	16.2%	35.4%	26.1%	4.6%	0.4%	0.0%	0.0%	0.0%
5	9.7%	0.0%	0.0%	0.0%	0.0%	7.5%	15.2%	35.8%	25.7%	5.6%	0.4%	0.1%	0.0%
6	6.6%	0.0%	0.0%	0.0%	0.0%	3.0%	6.9%	16.5%	34.2%	26.2%	6.0%	0.6%	0.0%
7	3.9%	0.0%	0.4%	0.0%	0.0%	2.0%	3.6%	6.4%	16.3%	35.2%	26.5%	5.6%	0.0%
8	2.5%	0.0%	5.2%	1.0%	0.0%	1.2%	2.2%	3.6%	7.2%	16.7%	34.0%	26.4%	0.0%
9	1.9%	0.0%	26.7%	5.9%	0.0%	0.7%	0.5%	2.0%	3.0%	7.5%	15.7%	35.0%	1.1%
12	0.1%	0.0%	2.0%	3.9%	0.0%	0.1%	0.2%	0.1%	0.3%	0.5%	0.7%	0.8%	91.2%

Row Percent
100.00
80.00
60.00
40.00
20.00
0.00

In the graphs, we can see the lift and Gain Charts for the ordinal logistic regression.





In the ordinal logistic regression as we have multiple ordered classes we have separate intercept and coefficient for each class, below we can see the parameter estimates for the ArrivalDelayGroup=0

Variable	Coefficient
Intercept	3.858
DepartureDelayGroups=0	0.682
DepartureDelayGroups=1	0.547
DepartureDelayGroups=10	-1.174
DepartureDelayGroups=11	-3.358
DepartureDelayGroups=12	-18.85
DepartureDelayGroups=2	0.445
DepartureDelayGroups=3	0.426
DepartureDelayGroups=4	0.421
DepartureDelayGroups=5	0.451
DepartureDelayGroups=6	0.389
DepartureDelayGroups=7	0.325
DepartureDelayGroups=8	0.19
dest_No_of_Run_ways	-0.002
DistanceGroup=1	-0.009
DistanceGroup=10	0.041
DistanceGroup=11	-0.004
DistanceGroup=2	0.008
DistanceGroup=3	0.009
DistanceGroup=4	0.009
DistanceGroup=5	0.016
DistanceGroup=6	0.021
DistanceGroup=7	0.017
DistanceGroup=8	0.036
LateAircraftDelay	-0.002
orig_No_of_Run_ways	0.001
Origin_visibility	0.002
Origin_wind_speed	-0.001

Variable	Coefficient
Intercept	2.535
DepartureDelayGroups=0	-4.129
DepartureDelayGroups=1	-4.088
DepartureDelayGroups=10	7.027
DepartureDelayGroups=11	27.634
DepartureDelayGroups=12	-14.6
DepartureDelayGroups=2	-3.987
DepartureDelayGroups=3	-3.894
DepartureDelayGroups=4	-3.817
DepartureDelayGroups=5	-3.577
DepartureDelayGroups=6	-3.606
DepartureDelayGroups=7	-2.83
DepartureDelayGroups=8	-2.791
dest_No_of_Run_ways	-0.007
DistanceGroup=1	0.058
DistanceGroup=10	0.184
DistanceGroup=11	0.023
DistanceGroup=2	0.111
DistanceGroup=3	0.116
DistanceGroup=4	0.055
DistanceGroup=5	0.093
DistanceGroup=6	0.144
DistanceGroup=7	0.115
DistanceGroup=8	0.132
LateAircraftDelay	-0.01
orig_No_of_Run_ways	0.004
Origin_temperature	0.002
Origin_wind_speed	-0.002

Below we can see the model for ArivalDelayGroup ‘0’ for the DepartreDelayGroup 12 and DistanceGroup 10:
3.8583 + (-18.8501)(DepartureDelayGroups_12) + (0.04119)(DistanceGroup_10) + (0.0011)((orig_No_of_Run_ways) + (-0.0080)(dest_No_of_Run_ways) + (0.0018)(Origin_temperature)+(-0.0012)(Origin_wind_speed)+(0.0020)(Origin_visibility)+(-0.00216)(LateAircraftDelay)
Below we can see the model for ArivalDelayGroup ‘1’ for the DepartreDelayGroup 1 and DistanceGroup 10:
1.3677 + (1.7577)(DepartureDelayGroups_1) + (0.0298)(DistanceGroup_10) + (-0.0043)((orig_No_of_Run_ways) + (-0.0019)(dest_No_of_Run_ways) + (0.0079)(Origin_temperature)+(- 0.0013)(Origin_wind_speed)+(0.0067)(Origin_visibility)+(-0.00631)(LateAircraftDelay)

VII. CONCLUSION AND RECOMMENDATIONS

By using the multinomial logistic regression model we are able to classify the arrival delay of an incoming aircraft with around 85% accuracy, and if you observe the classification table we are able the classify the two extreme groups namely arrival delay group '0' and '12' with more than 95% accuracy but the remaining groups accuracy is low compared to these groups, this is revealing that the groups are very close to each other, So, by combining few of the groups into one or recreating the groups with sufficient distance between the groups may increase the accuracy of the classification.

BIBLIOGRAPHY

[1] ALAN AGRESTI, Categorical Data Analysis, Second Edition, John Wiley & Sons, Inc.,
[2] ANN A. O'Connell, LOGISTIC REGRESSION MODELS FOR ORDINAL RESPONSE VARIABLES, SAGE PUBLICATIONS
[3] DAWID W. HOSMER, Applied Logistic Regression, Second Edition, John Wiley & Sons, Inc.,
[4] John O. Rawlings, Sastry G. Pantula, David A. Dickey, Applied Regression Analysis: A Research Tool, Second Edition, Springer
[5] Damodar Gujarati, Econometrics by Example, PALGRAVE MACMILLAN
[6] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer