# Flight Delay Classification


## Ritik Bilala


GNR 652
Assignment 2

## INTRODUCTION

Problem is a classification problem , solved using Logistic Regression in this assignment.

Data include 13 Columns with Numerical as well as categorical features

Numpy , Pandas and scikit_learn libraries are used

## FEATURES GIVEN IN TABLE

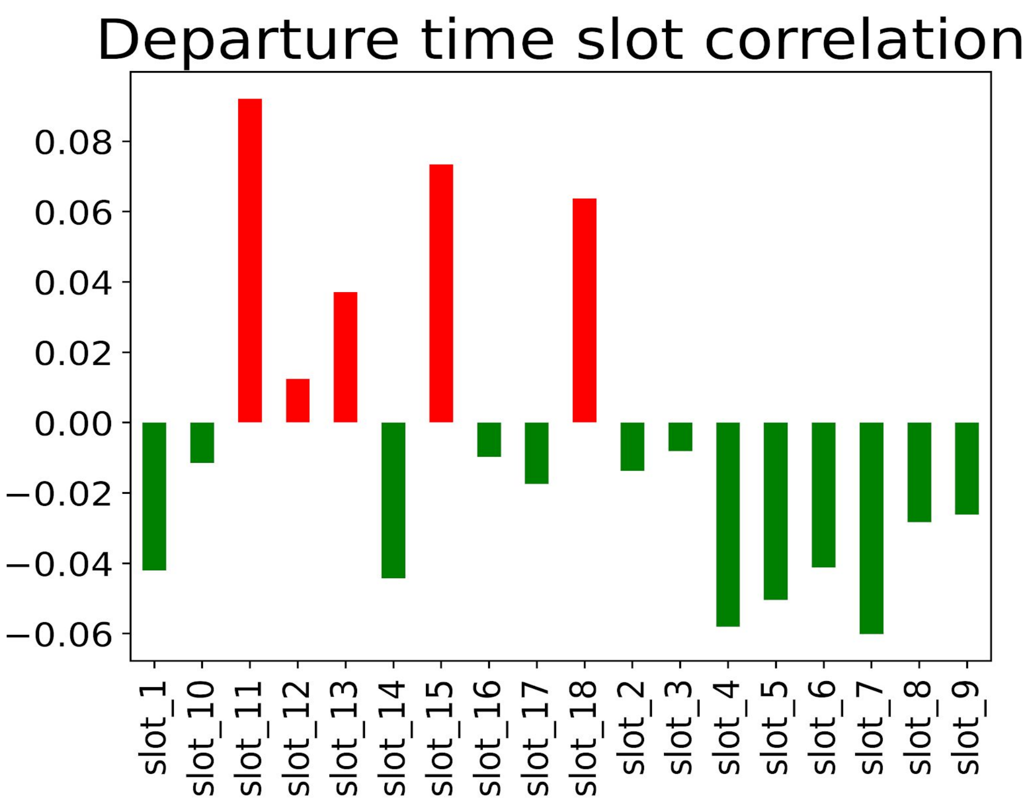
1. CRS
2. Actual Departure: Delay in departure = Actual in Departure - CRS
3. CARRIER
4. ORIGIN
5. DESTINATION
6. Date (dd/mm/yy)
7. Flight Number
8. Tail Number
9. Weather
10. Day of week
11. Day of month
12. Distance

## HYPOTHESIS

1. One by one analysis of every feature and making relevant assumptions regarding features
2. As most of the data is categorical , other numerical datas are converted into categories by making relevant grouping among them to ease the model
3. ONE_HOT_ENCODING of variables is done to create dummy variables and finding correlation of each category with delaying of flight
4. Dichotomous variables are kept as 1 or 0 (no OneHotEncoding)

## CRS_TIME

1. CRS_TIME is a numerical data which is converted into categories of 18 slots between 6 am to 10pm.
2. Slot_11 is has highest correlation to delay of flight , still values are small



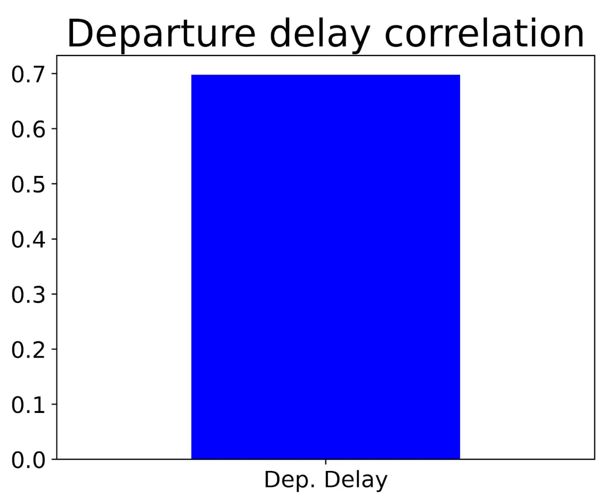
## DEPARTURE TIME

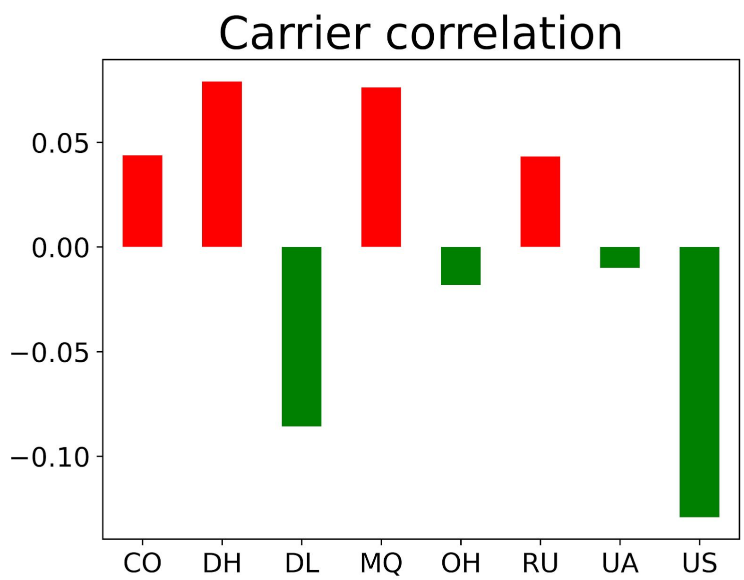
Instead of using departure time directly , its difference from reporting time is used, and is more relevant as delays in departure can lead to delay in flight journey

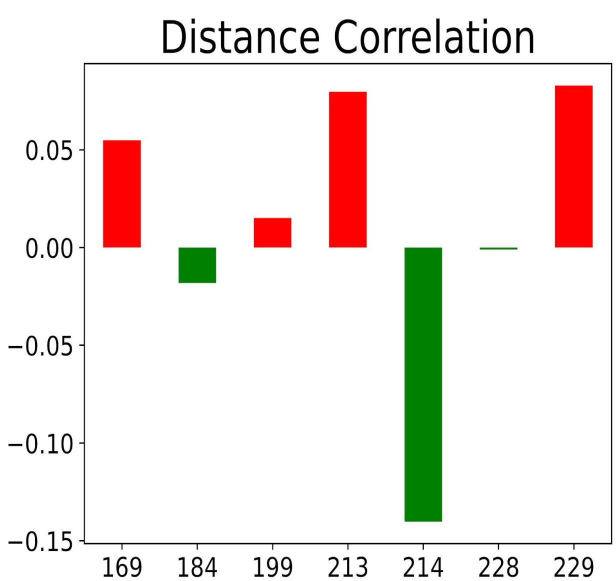
****The correlation of departure delay with flight getting delayed is high and was expected**

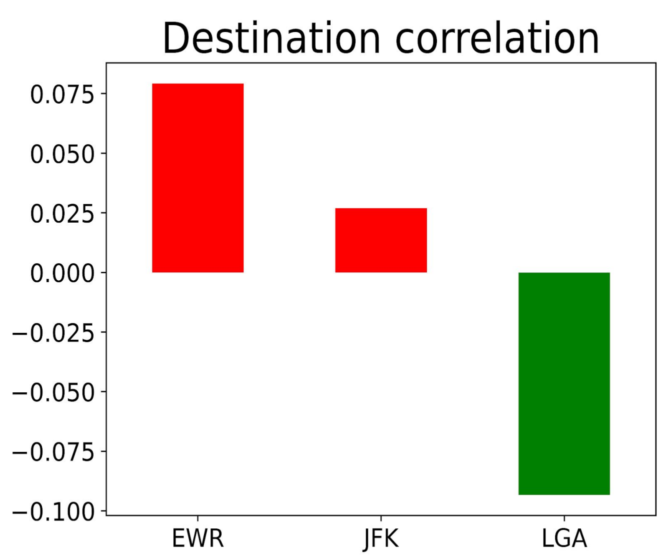
## CARRIER

1. **US** has a strong negative correlation which means it is associated with being 'ontime'. Value of correlation is weak for others



## Destination ,and Origin

'LGA' destination  and Origin 'DCA' Is highly correlated with being on time . although correlation are weak for both

## Distance

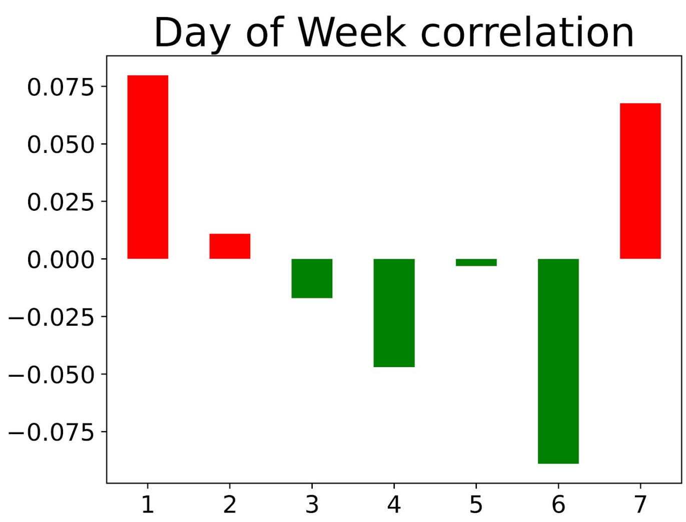
Value 214 is highly correlated with being on time. But from data distance is unique  for given 'origin and 'destination' pair and we can remove this feature by keeping other two

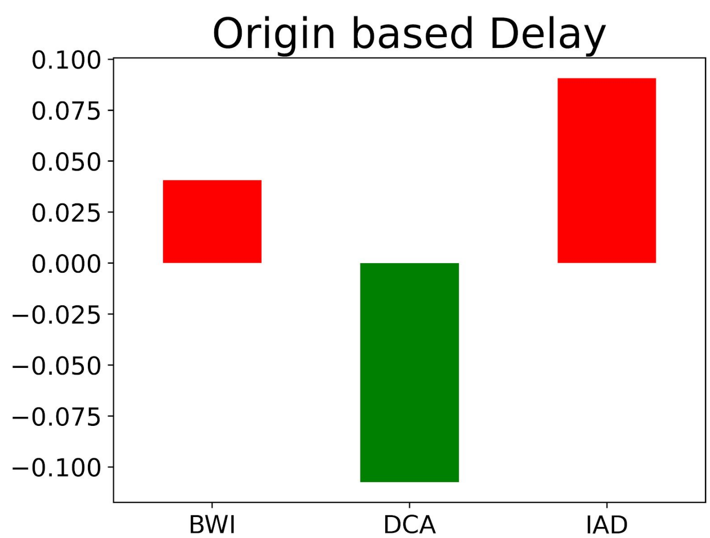
## Date, Day of week and Day of month

Data is given only for January month hence month No. cannot be a feature.

Day of week seems more reasonable as it repeats over a cycle , day of month can be remove while keeping day of week feature

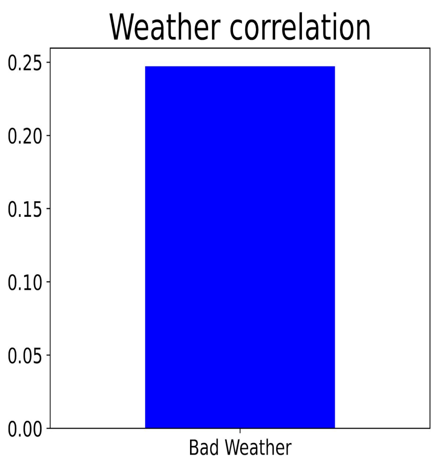
Day 6 or Saturday is correlated with being ontime, the correlation is weak



## Weather

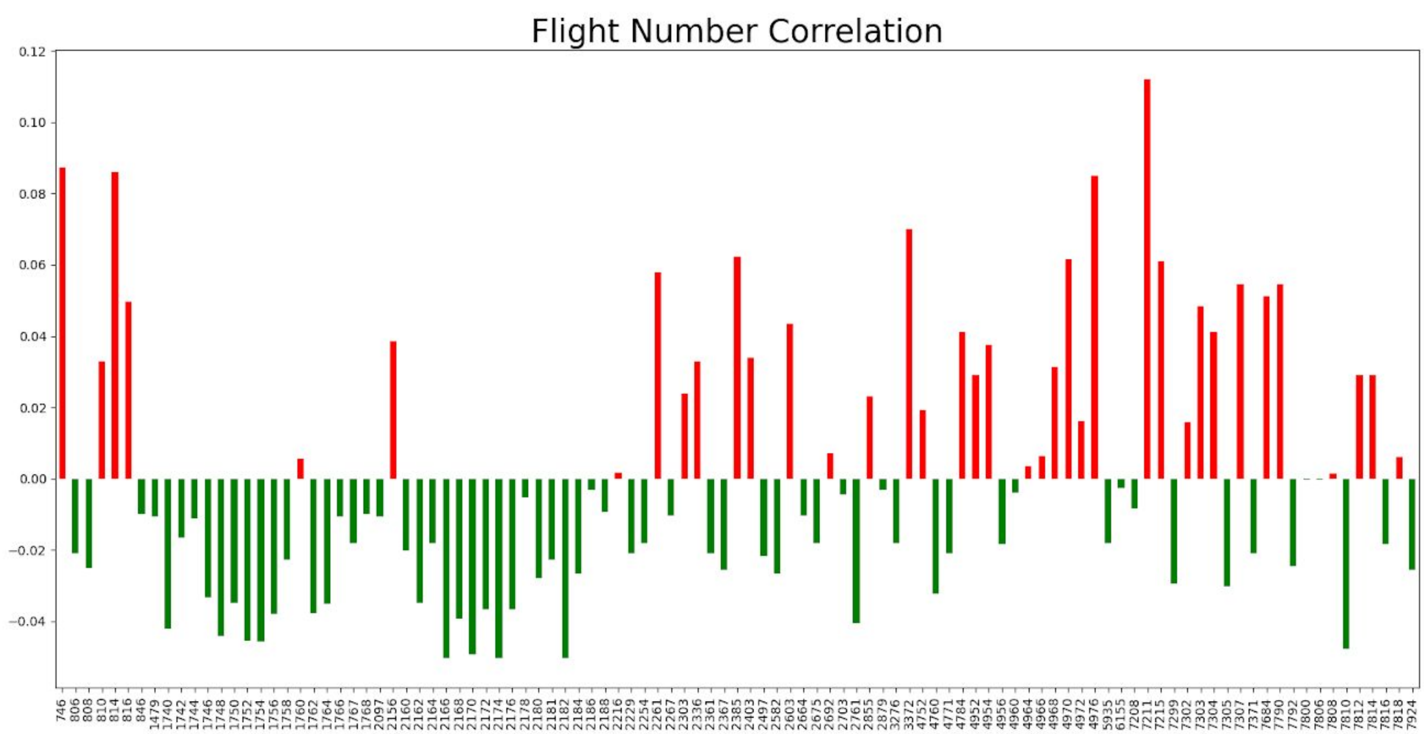
BAD weather is highly correlated with flight being delayed




## Flight Number and Tail number

Some flight numbers such as 7211 are highly correlated with being on time .

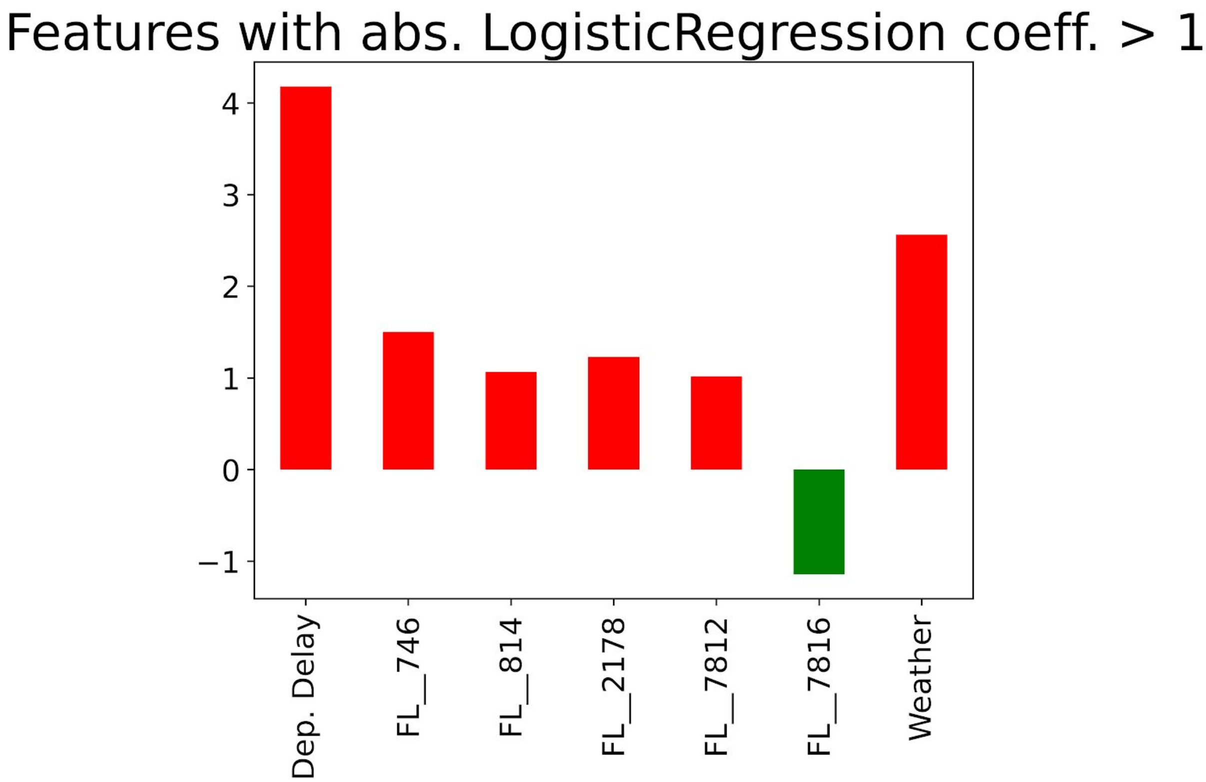
Tail number is unique for given carrier and destination , hence can be removed



## LOGISTIC REGRESSION MODEL WITH ALL FEATURES ( 9 selected on basis of arguments)

### Accuracy : 91.94

Every feature is categorised and dummy variables were introduced for every category. For departure delay two categories were introduced delay >15 min or <15 min. For CSR 18 time slots were used as categories




**Analysis : The correlations of Departure delay, weather and flight number were found to be high and the logistic regression coefficients found consistent to our expectations.**

**Conclusion:**

1. Bad weather and departure delays are strong predictor of actual delay
2. Some of the flight numbers are strong predictor for delay or ontime but rest are useless
3. This way 151 variables were used which can make prediction less accurate

## FEATURE SELECTION:

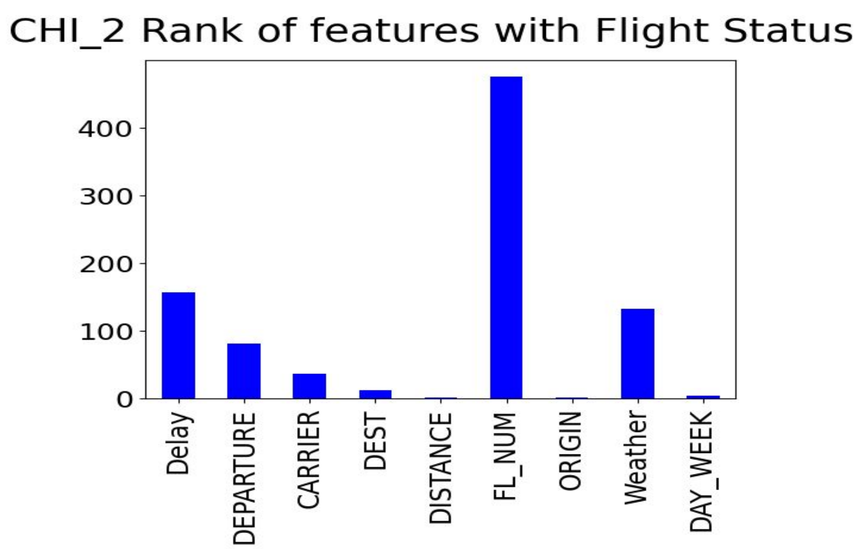
### METHOD 1

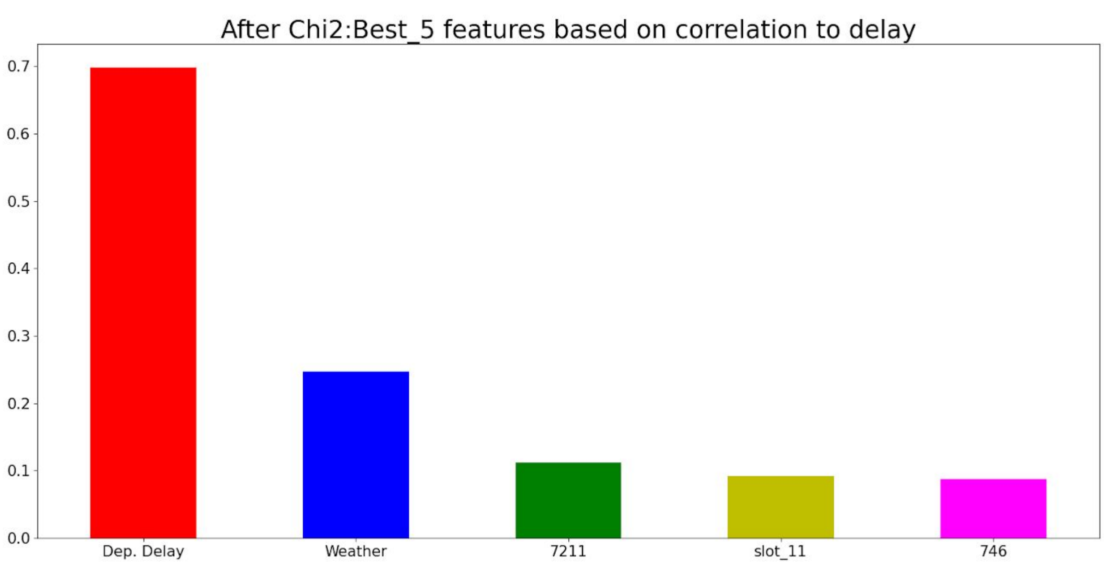
## Accuracy : 92.05

1. DATE, TAIL NUMBER, DAY OF MONTH, DEPARTURE TIME were removed as they are related with other features in the data (argument is given with description of every feature)
2. **CHI- SQUARE** test was done to find **BEST 4 given categorical features**
3. Dummy variables were introduced for BEST 4 features and top 5 dummy variables were selected on basis of absolute value of correlations to flight status

FL_NUMBER , WEATHER, DEP. DELAY, CSR time

| Feature | Model coefficient |
|---|---|
| Departure Delay | **3.97524005** |
| Bad Weather | **0.44249852** |
| FL_7211 | **1.64149737** |
| SLOT_11 | **0.06529826** |
| 746 | **2.5813288** |



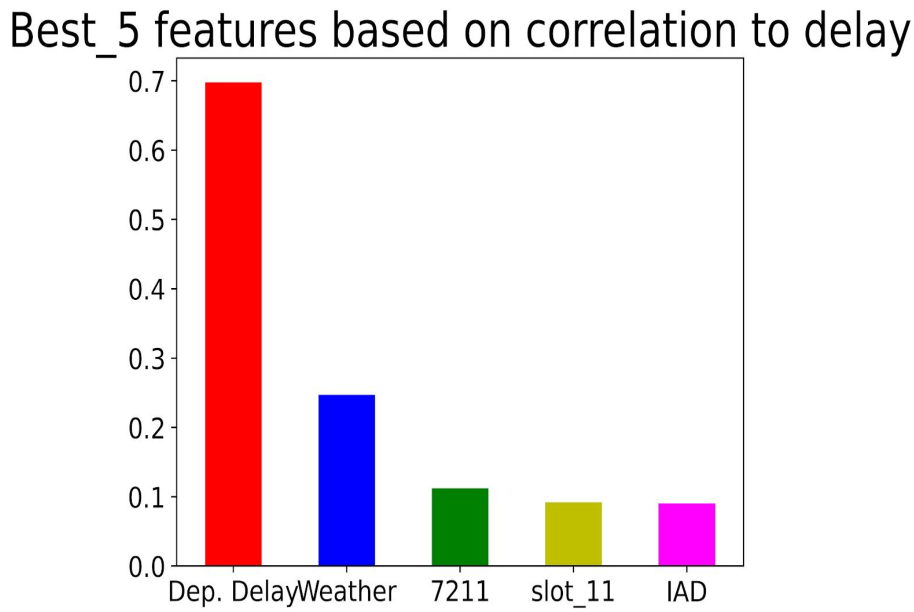



**METHOD 2:**

## Accuracy : 92.05

1. Dummy variables were introduced for all 9 variables and best 5 dummy variables were selected on basis of absolute value of correlation with flight status

| Feature | correlation | Model coefficient |
|---|---|---|
| Departure Delay | 0.697911 | **3.93926803** |
| Bad Weather | .247217 | **0.69708248** |
| FL_7211 | .112191 | **-0.09917736** |
| SLOT_11 | .092112 | **0.0315907** |
| IAD | .090716 | **2.63390981** |




## RESULTS

Model coefficients of logistic regression with all features and accuracy after feature selection by both methods shows that:

1. Bad weather
2. Departure delay
3. Flight number
4. Departure slot

Are important feature to be considered in prediction

*****Certain flight number (7211) and departure slots (slot_11) have strong correlation with the flight status and rest other categories can be neglected for reducing dimensions and increasing accuracy**

## QUESTIONS

**Question 6** : Find the ideal weather conditions for the highest chance of an ontime flight from DC to New York (weather, time, day, carrier)

**Answe**r :From correlation table given with description of features :

**Weather: good weather**

**Time: Slot 7 ie: 12:00 to 12:45 pm**

**Day: Saturday**

**Carrier: US**

**Are the best conditions for flight being on time**

## BONUS:

Q1 :

VERONICA, ULTRON , KAREN

Q2:

The Data processing inequality is an information theoretic concept which states that the information content of a signal cannot be increased via a local physical operation. This can be expressed concisely as 'post-processing cannot increase information'.

Q3:

The Rule Of Two

Q4:

C-3PO and R2-D2

Q5:

AI algorithms write cards and compete with writers