ASSIGNMENT 02 REPORT

Q1) For the visualization of the dataset I have used the following methods

Bar Graph:

a) To plot the number of various carriers, origin and destination cities using the value_counts function.
b) To find the number of flights that are on time and delayed in total dataset using the value_counts function.
c) Also, plot the different features (carrier, day_week, weather, origin, destination) with respect to flight status using count plot from sns library.

KDE(kernel density estimation):

a) I used this KDE to plot the features (distance and dept_time) which have continuous values.

Pie Chart:

a) To get more information about the carriers, how these different carriers are sharing the percentage of the total for flight status.

Heatmap:

a) I used heatmap to find out the correlation between the numerical features which are mentioned in the dataset. What I found is that the correlation coefficient between the dept_time and crs_dept_time, weather and day_week are higher compare to other features.

Q2) In Preprocessing of the dataset first I standardized the dataset since there are three features that have high range values and did the same using sklearn package. Dropped the unnecessary features like FL_NUM, TAIL_NUM, FL_DATE.
I had converted the flight status using label_encoder to ontime=1 and delay=0.
From sklearn I have used library and converted all the category features into dummy variables.

Q3) After separating the dataset and fitting the test dataset onto the model I got the following accuracies.

a) 80.9 % accuracy using a logistic regression model from scratch (that is defining loss function, finding its gradient, and then updating weights).
This model is not perfect and for some of the values, it is misclassifying the data.
b) 87% accuracy using a logistic regression model from sklearn. And other parameters from the model I got are following,

Accuracy of Logistic regression classifier on test set: 0.87

	precision	recall	f1-score	support
0	1.00	0.29	0.45	163
1	0.86	1.00	0.93	718
micro avg	0.87	0.87	0.87	881
macro avg	0.93	0.64	0.69	881
Weighted avg	0.89	0.87	0.84	881

Confusion matrix:

Coefficients obtained from the model:

4.2742655 , -4.70931008, 0.02129813, -0.76449997, 0.08938631, -0.16615835, -0.18373648, 0.01734126, 0.14790424, -0.21716511, 0.1305248 , -0.18994231, 0.02383552, 0.26620417, 0.0396631 , 0.05142842, -0.07651556, -0.00573257, 0.08592897, -0.08635132

Value of the intercept:

= 1.5585

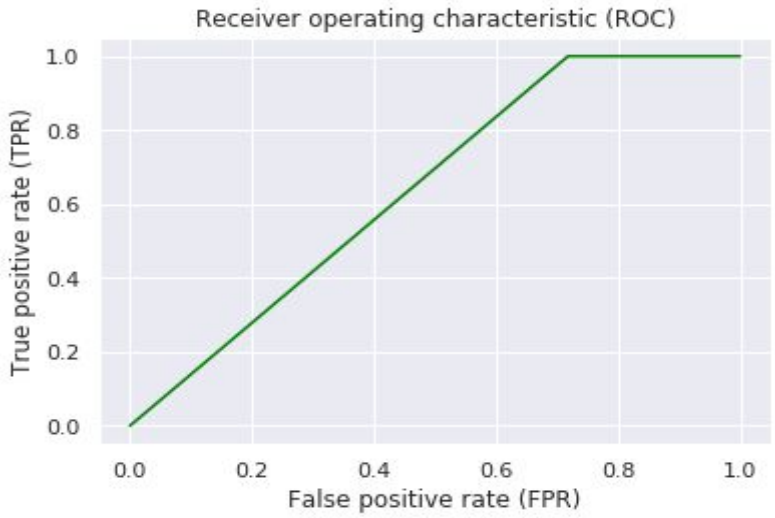
Measures of model performance on test set:

RMSE = 0.36442

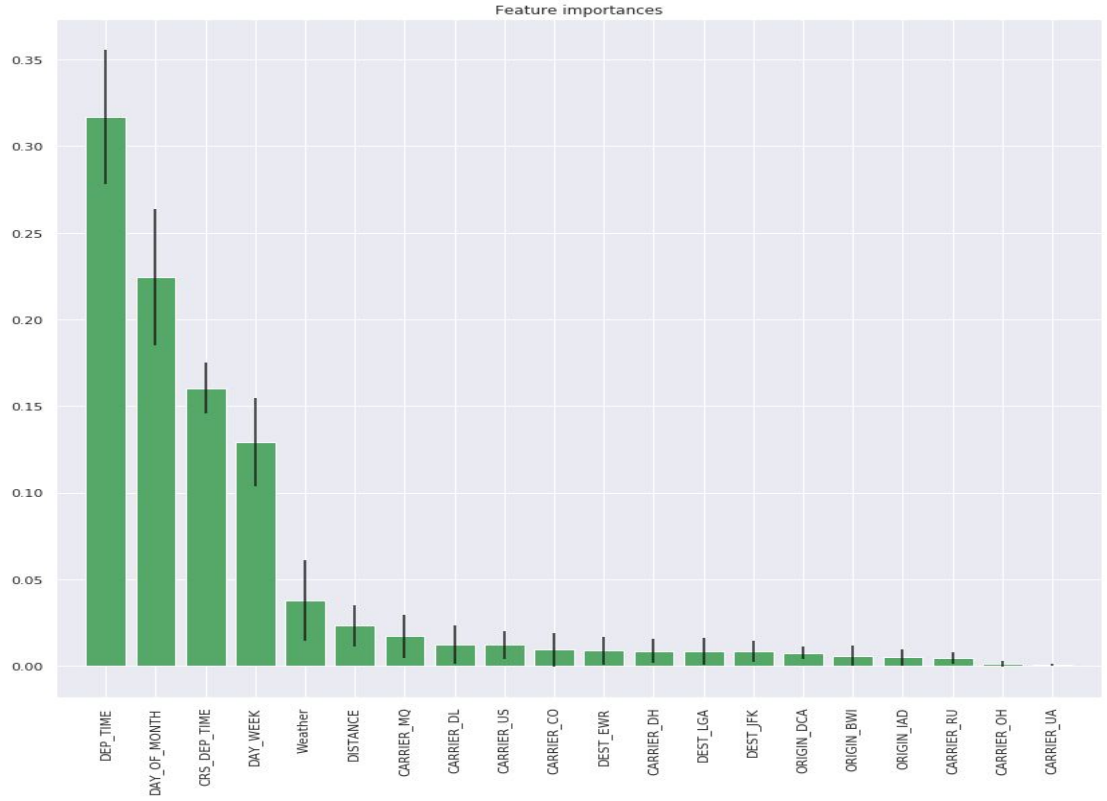
R_Squared = 0.11925

ROC Curve:

The area under the ROC curve: 0.3588



Q4) Performed variables selection using tree-based feature selection (RFE). On the basis of the feature importance of the forest, I plotted the graph between features and feature importance.



I chose the starting 8 features for my model, and further analyzed the model results on fitting the same logistic regression model on the extracted features.

Features Important are:

dep_time, day_of_month, crs_dep_time, day_week, weather, distance, carrier_mq, carrier_dl.

And the remaining features are not much of importance.

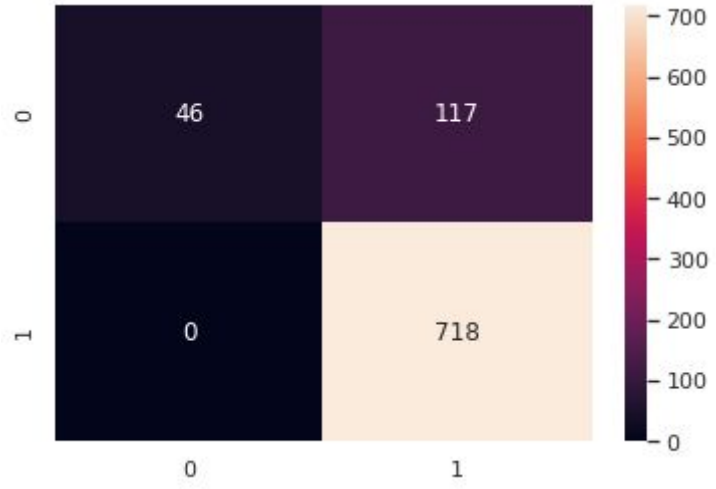
Q5) After fitting the model on the extracted features I got the following results,

a) Accuracy of the same is 87%

Accuracy of Logistic regression classifier on test set: 0.87

	precision	recall	f1-score	support
0	1.00	0.28	0.44	163
1	0.86	1.00	0.92	718
micro avg	0.87	0.87	0.87	881
macro avg	0.93	0.64	0.68	881
Weighted avg	0.89	0.87	0.84	881

Confusion matrix:



Coefficients :

4.58058436, -5.01615761, 0.07697328, -0.77150093, 0.08834864, -0.15181712, 0.18530203, -0.16382605

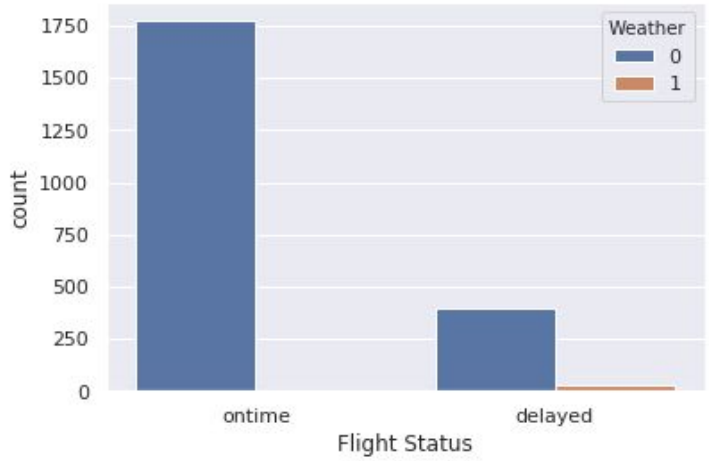
Intercept:

= 1.497

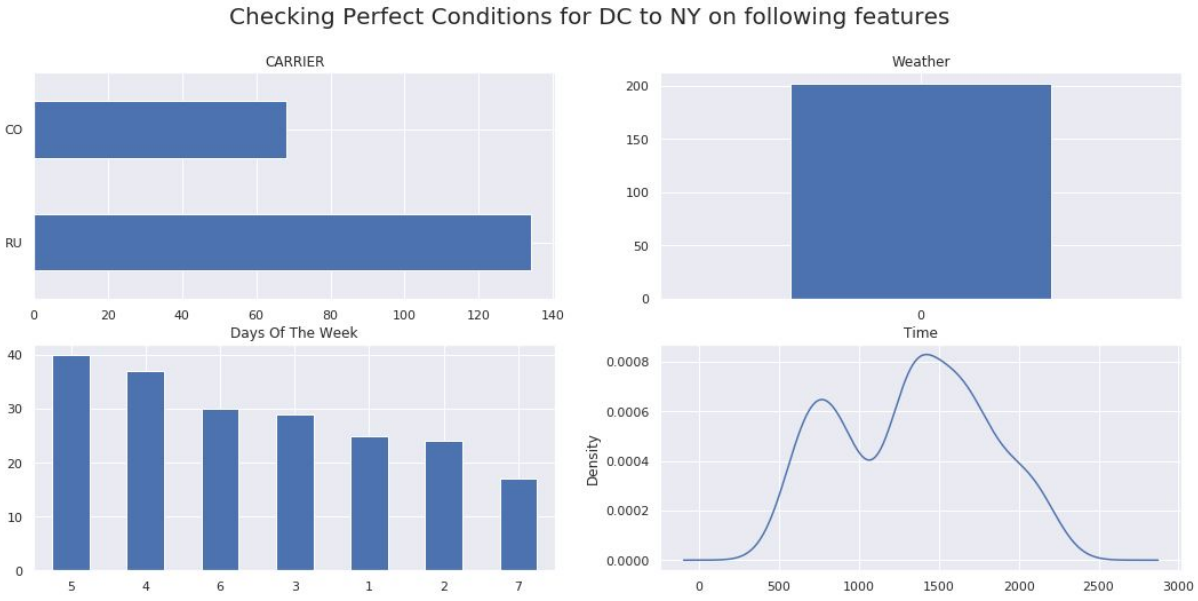
Performance measures on test set:

RMSE= 0.3644 and R_Squared = 0.1192

Q6) Ideal weather conditions for the highest chance of an on-time flight from DC to New York.



Analytically from the plot, we can see that when the flight is on-time that time weather is always 0.



But after extracting the dataset according to question and plotting it we got the following perfect conditions on given features.

Weather = weather condition 0

Time = around 1500

Day_week = 5th

Carrier = RU.

BONUS

1) Name any AIs made by Tony Stark in the Marvel Cinematic Universe besides JARVIS, FRIDAY, and EDITH.
    - P.L.A.T.O. also called Piezo-electrical Logistic Analytical Tactical Operator is created using the same kind of programming of H.O.M.E.R. mainly for the works in Stark industries and Maria Stark Foundation. P.L.A.T.O. was also given a 3-d holographic body, to help him communicate with the employees of Stark Industries. When Stark reformed the Avengers West Coast into Force Works, he moved them into the Works facility and P.L.A.T.O. helped the team in various ways. It's assumed that P.L.A.T.O., like H.O.M.E.R., was disassembled when Tony was temporarily believed to be
    - Others AIs are *Heuristically Operative Matrix Emulation Rostrum(* H.O.M.E.R.*), Virtual Integrated Rapidly-evolving Grid-based Intelligent Lifeform(V.I.R.G.I.L)*

2) Explain the Data processing inequality
    - Intuitively, the data processing inequality says that no clever transformation of the received code (channel output)Ycan give more information about the sent code (channel input)X than Y itself

      (Data processing inequality)
      Suppose we have a probability model described by the following (Markov Chain): $X \rightarrow Y \rightarrow Z$ where $X \perp (Z|Y)$, then it must be that $I(X, Y) \geq I(X, Z)$

3) X is a **Rule Of Two.**

4) **C-3PO** and **R2-D2** pronounced **Artoo-Detoo** and often referred to as **R2 (Artoo)**.

5) The specialty about the Cards against Humanity is, it was a computer algorithm designed to compete with humans for 16 hours to find out which one has the best pack of cards.