**@Class  Here is the rubric I will be using to grade Lab 1.  You must have these exact sections in your project, if you want me to grade it!  Lab 1 is due Sunday, 05/16/2021 by 11:59PM.**

|  |  |  |
| --- | --- | --- |
| Category | Points | Description |
|  |  |  |
| Business Understanding | 10 | Describe the purpose of the data set you selected (i.e., why was this data collected in the first place?). Describe how you would define and measure the outcomes from the dataset. That is, why is this data important and how do you know if you have mined useful knowledge from the dataset? How would you measure the effectiveness of a good prediction algorithm? Be specific. |
| Data Meaning Type | 10 | Describe the meaning and type of data (scale, values, etc.) for each attribute in the data file. |
| Data Quality | 15 | Verify data quality: Explain any missing values, duplicate data, and outliers. Are those mistakes? How do you deal with these problems? Give justifications for your methods. |
| Simple Statistics | 10 | Visualize appropriate statistics (e.g., range, mode, mean, median, variance, counts) for a subset of attributes. Describe anything meaningful you found from this or if you found something potentially interesting. Note: You can also use data from other sources for comparison. Explain why the statistics run are meaningful. |
| Visualize Attributes | 15 | Visualize the most interesting attributes (at least 5 attributes, your opinion on what is interesting). Important: Interpret the implications for each visualization. Explain for each attribute why the chosen visualization is appropriate. |
| Explore Joint Attributes | 15 | Visualize relationships between attributes: Look at the attributes via scatter plots, correlation, cross-tabulation, group-wise averages, etc. as appropriate. Explain any interesting relationships. |
| Explore Attributes and Class | 10 | Identify and explain interesting relationships between features and the class you are trying to predict (i.e., relationships with variables and the target classification). |
| New Features | 5 | Are there other features that could be added to the data or created from existing features? Which ones? |
| Exceptional Work | 10 | You have free reign to provide additional analyses. One idea: implement dimensionality reduction, then visualize and interpret the results. |
|  |  |  |
| Total | 100 |  |

**Business Understanding**

Describe the purpose of the data set you selected (i.e., why was this data collected in the first place?). Describe how you would define and measure the outcomes from the dataset. That is, why is this data important and how do you know if you have mined useful knowledge from the dataset? How would you measure the effectiveness of a good prediction algorithm? Be specific.

The data set for this Lab Project is the Airline Passenger Dataset sourced from Kaggle (<https://www.kaggle.com/binaryjoker/airline-passenger-satisfaction>). This dataset comes from a survey that measured Airline passenger satisfaction and to understand the various factors that contribute to airline passenger perceived level of satisfaction. The survey was conducted to gather data so the airline companies could improve on Quality of Service (QoS) that will meet the expectations of passengers and help these companies stay competitive within the industry. The model when built will help these companies direct resources that will target specific attributes to improve on services hence saving them millions of dollars that would otherwise gone to a gamut of service improvement campaigns. The dataset can be leveraged into a supervised machine learning classification model that can predict customer satisfaction for this dataset set as well as predict future flights experiences based on attributes and datapoints used.

Statistical models that can be used for the desired outcome for this analysis include PCA as a dimensionality reduction method to reduce the dimensionality of the variables in this large data sets into a smaller ones that still contains most of the information we will need. Logistic regression, random forests, naive-bayes, KNN or LDA/QDA will also be employed. The expected outcome of such a model after dimensionality reduction would be to correctly predict a passenger’s flight satisfaction description from training/test set split or k-fold cross validation while maintaining high and consistent measurement for overall model accuracy, sensitivity and specificity.

Cross validation steps will ensure that the model was not overfitted based on training data and will work for future data that has either not yet been collected or where all other attributes are known but not satisfaction. By maximizing both sensitivity and specificity the model is robust to a skewed sample or dataset, the amount of false positives is just as important as the ones that were matched correctly. In addition, By running several classification model types we can ensure that multi approaches were taken and tested out before a model its accuracy statistics are deemed final.

**Data Meaning Type**

Describe the meaning and type of data (scale, values, etc.) for each attribute in the data file.

The summary statistics of the Airline data consist of 129880 records with 24 attributes. These attributes are both continuous and categorical variables with 393 missing values as seen below. The variables are Unnamed, Gender, Customer\_type, Age, Type\_of\_travel, customer\_class, flight\_distance, inflight\_wifi\_service, departure\_arrival\_time\_convenient, ease\_of\_online\_booking, gate\_location, food\_and\_drink, online\_boarding, seat\_comfort, Inflight\_entertainment, Onboard\_service, leg\_room\_service, baggage\_handling, checkin\_service, inflight\_service, cleanliness, departure\_delay\_in\_minutes, arrival\_delay\_in\_minutes, satisfaction.

A detail descriptive summary statistics below shows the number of entities of the data, the means, standard deviations, each variable’s minimum and maximum values as well as their interquartile percentiles giving us a fair idea of the cardinality of the data’s distribution and spread. These values arouse curiosity to further investigate if they violates any statistical model’s assumptions or require further transformation that will help build an accurate predictive model. A scatter matrix plot will give us a better pictorial view of the data’s distribution.

Data Quality

Verify data quality: Explain any missing values, duplicate data, and outliers. Are those mistakes? How do you deal with these problems? Give justifications for your methods.

A careful look at the data’s statistical plots, we can see the some variables have outliers and missing values.

For the 393 missing values which constitute about 0.3% of the total arrival\_delay\_in\_minutes count, we are going to impute that variable’s missing values with values from the next row to fill-in using the pandas function **(airline.interpolate(method='linear', direction = 'forward', inplace=True))** in python to help us minimize errors. The confirmation print below hence shows a clean dataset to enable us proceed with confidence.

The **“unnamed**” variable row will be dropped from the dataset as there was we aren’t certain if there was an error in capturing the data for that unknown attribute. If included, we cannot make any meaningful interpretation with it hence the approach to drop it entirely. This reduces the total predictors to 23. This we believe will enhance the quality of data to build the model.

Outliers:

Using the 68-95-99 rule, we can detect outliers in the various attribute. To explain this rule 68% of the data is within one standard deviation above or below the mean, and 95% of the data is within two standard deviations from the mean, also 99.7% of the data is within three standard deviations from the mean. With this rule of thumb only very few data points should be beyond three standard deviations from the mean, more precisely, only 0.3% of the data points. Therefore, any data point that is seen farther than three standard deviations is considered extreme and an outlier.

To check if a data point is an outlier and check if it falls farther than three standard deviations, we calculate:

Q1– (1.5 \* IQR),

Q3 + (1.5 \* IQR).

These represent the lower and upper bounds of the area in the distribution that is not considered extreme. Which ends up being approximately 3 standard deviations from the mean. Hence, any data point lower than the lower bound or greater than the upper bound is an outlier and they are greater than the 3 standard deviations. Implying;

(ant data point value) < Q1– (1.5 \* IQR), then it’s an outlier.

(any data point value) > Q3 + (1.5 \* IQR), then it’s an outlier

That said, checking the data for outliers shows that the attributes **'checkin\_service'**, **‘flight\_distance’**, **‘departure\_delay\_in\_minutes’** and **‘arrival\_delay\_in\_minutes’** have outliers in then that require further investigation.

For fewer outliers within the **'checkin\_service'** variable, we can delete them as there are not significant and those of **‘flight\_distance’**, **‘departure\_delay\_in\_minutes’**, **‘arrival\_delay\_in\_minutes’** the team will transform the datapoints to normalize the distribution.

After these processes have been done, a correlation plot will be plotted to check for variable correlation in the dataset to measure their influence on the outcome.

**Simple Statistics**

Visualize appropriate statistics (e.g., range, mode, mean, median, variance, counts) for a subset of attributes. Describe anything meaningful you found from this or if you found something potentially interesting. Note: You can also use data from other sources for comparison. Explain why the statistics run are meaningful.

A boxplot with attributes of interest with outliers was plotted as seen below to see how they stack out. The attributes were **'checkin\_service', ‘flight\_distance’, ‘departure\_delay\_in\_minutes’,** and **‘arrival\_delay\_in\_minutes’.** The boxplot shows that **‘departure\_delay\_in\_minutes’,** and **‘arrival\_delay\_in\_minutes’** are heavily skewed to the left suggesting a right skew will be transformed and that of **'checkin\_service'**  is left tailed skew with two outliers which can be removed whereas the **‘flight\_distance’** attribute is right tailed with some sizable amount of outliers can be transformed.

A pairplot of the above mentioned attributes shows the scatterplot distribution of spread or variance by each response in the satisfaction category. The scatterplot indeed shows a pairplot of what we already know about the outliers but narrows it down to which responses exhibits these outliers.

Measure of Central tendencies:

The measure of central tendencies from our EDA shows that, the mean age in the dataset is 39 years and with the median being 40 years. The age’s 25th percentile is 27 years and its 75th percentile is 51 years. The age range is 24 years with a standard deviation of 15.

The checkin \_service shows a mean value of 3.3 a 0.3 units above the 25th percentile of 3.0. the 75th percentile is 4 and a median of 3. The interquartile range for checkin\_service is 1 with its lower whisker and upper whisker being 1.5 and 5.5 respectively. The attribute’s standard deviation is 1.26

The mean for this attribute is four that of flight\_distance is 1190 which is slightly above median value of 844 with 346 points. The 25th and 75th percentiles are 414 and 1744 respectively.

The departure\_delay\_in\_minutes attribute has a mean value of 14.71 minutes and a median value of 0. The 25th and the 75th percentiles are 0 ad 12 minutes having an interquartile range of 12 minutes. The standard deviation here is 38.07.

The arrival\_delay\_in\_minutes attribute also has a mean value of 150.9 minutes with a median value of 0 and a standard deviation of 38.43. The 25th and 75th percentiles are 0 and 13 minutes with a range of 13.

Explain why the statistics run are meaningful.??

To be discussed

**Visualize Attributes:**

Visualize the most interesting attributes (at least 5 attributes, your opinion on what is interesting). Important: Interpret the implications for each visualization. Explain for each attribute why the chosen visualization is appropriate.