Predictive ML Model for Airline Success

Tiwari Ayushi
Data Analytics and Engineering Dept.

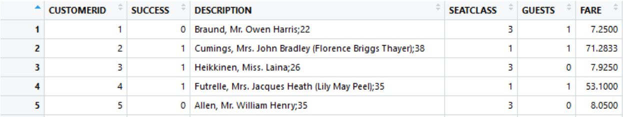
Volgenau School of Engineering
Herndon, USA
atiwari4@masonlive.gmu.edu

Abstract— **As a data scientist for airline A, analyze a customer database to identify the factors that clarify why some customers are flying with it, while others are canceling. The data science team wants to identify the important factors and help the advertising team develop customer-demographic-specific packages to attract more customers. The key task here is to extract metadata and use it to create a classification/prediction model for the following problem: given a customer, will he or she fly or not fly with Airline A?**

Keywords—descriptive, metadata, python, DBpedia Spotlight

I. Introduction

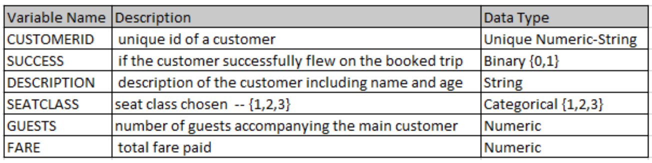
The raw dataset containing the airline customers information was imported in JSON and semi-structured format. It was converted to comma separated values using online conversion tool. After some initial column reordering, the dataset was imported in R Studio for analysis, manipulation and missing values imputation.



	CUSTOMERID	SUCCESS	DESCRIPTION	SEATCLASS	GUESTS	FARE
1	1	0	Braund, Mr. Owen Harris;22	3	1	7.2500
2	2	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer);38	1	1	71.2833
3	3	1	Heikkinen, Miss. Laina;26	3	0	7.9250
4	4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel);35	1	1	53.1000
5	5	0	Allen, Mr. William Henry;35	3	0	8.0500

Fig 1 Original Dataset

The original dataset has 891observations and 6 variables. The variables are CUSTOMERID, SUCCESS, DESCRIPTION, SEATCLASS, GUESTS and FARE. The description of variables is given below –



Variable Name	Description	Data Type
CUSTOMERID	unique id of a customer	Unique Numeric-String
SUCCESS	if the customer successfully flew on the booked trip	Binary {0,1}
DESCRIPTION	description of the customer including name and age	String
SEATCLASS	seat class chosen -- {1,2,3}	Categorical {1,2,3}
GUESTS	number of guests accompanying the main customer	Numeric
FARE	total fare paid	Numeric

Fig 2 Dataset Description

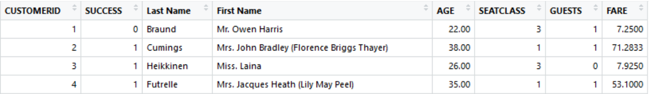
The objective is to first pre-process the original dataset, then extract descriptive metadata (age and gender) from variable DESCRIPTION and perform exploratory data analysis after performing missing values imputation.

In the later part of the project we will apply Machine Learning algorithms to predict who will fly with Airline A.

II. Descriptive Metadata extraction

DBpedia Spotlight Demo was used for descriptive metadata extraction.

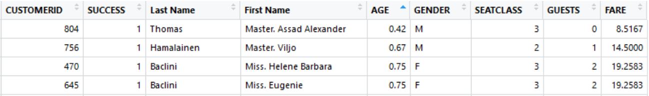
The task includes to extract descriptive metadata (age and gender) from variable DESCRIPTION and perform exploratory data analysis after performing missing values imputation. When we extract descriptive metadata from variable DESCRIPTION, we observe missing values in variable AGE. The dataset was imported in R Studio after extraction. The new dataset looks as below –



CUSTOMERID	SUCCESS	Last Name	First Name	AGE	SEATCLASS	GUESTS	FARE
1	0	Braund	Mr. Owen Harris	22.00	3	1	7.2500
2	1	Cumings	Mrs. John Bradley (Florence Briggs Thayer)	38.00	1	1	71.2833
3	1	Heikkinen	Miss. Laina	26.00	3	0	7.9250
4	1	Futrelle	Mrs. Jacques Heath (Lily May Peel)	35.00	1	1	53.1000

Fig 3 Dataset after metadata (AM_1) extraction

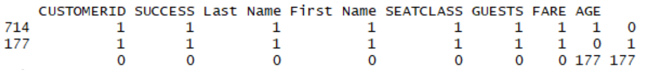
Below is the dataset sample after (GENDER) AM_2 was extracted.



CUSTOMERID	SUCCESS	Last Name	First Name	AGE	GENDER	SEATCLASS	GUESTS	FARE
804	1	Thomas	Master. Assad Alexander	0.42	M	3	0	8.5167
756	1	Hamalainen	Master. Viljo	0.67	M	2	1	14.5000
470	1	Baclini	Miss. Helene Barbara	0.75	F	3	2	19.2583
645	1	Baclini	Miss. Eugenie	0.75	F	3	2	19.2583

Fig 4 Dataset after metadata (AM_2) extraction

A. *Missing values imputation*



```
    CUSTOMERID SUCCESS Last Name First Name SEATCLASS GUESTS FARE AGE
714          1       1         1          1         1      1    1   0
177          1       1         1          1         1      1    0   1
             0       0         0          0         0      0    0 177 177
```

The above denotes there are 177 cases where value of AGE for the customer is missing.

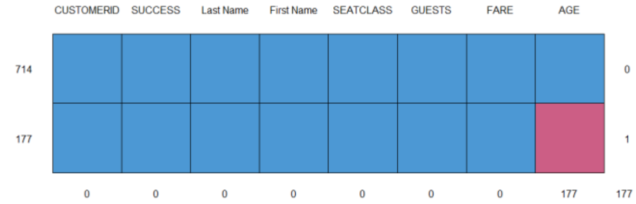


Fig 5 Missing data

The above figure shows that AGE variable has missing values.

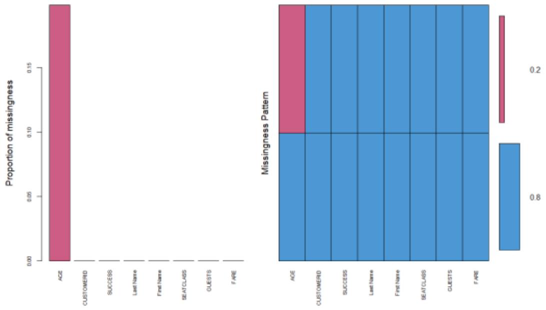


Fig 6 Proportion of missing values

We see that the AGE variable has 20% missing values.It also shows the different types of missing patterns and their ratios. The next thing is to draw a margin plot which is also part of VIM package.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

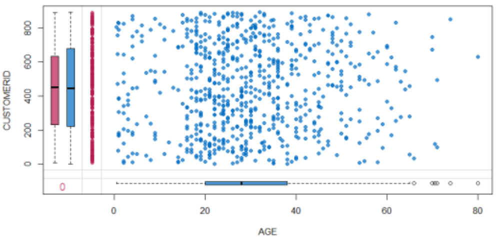
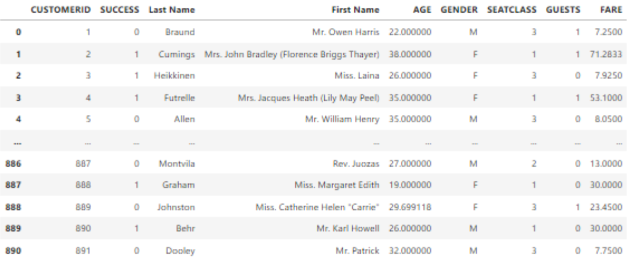


Fig 7 Margin Plot for missing data

The margin plot, plots two features at a time. The red plot indicates distribution of one feature when it is missing while the blue box is the distribution of all others when the feature is present. This plot is useful to understand if the missing values are Not Missing at Random (NMAR). For NMAR values, the red and blue boxes will be identical.

We applied mean values imputation method to fill missing values in AGE variable. Through this method the mean of the AGE variable remains undisturbed. However, the skewness of data enhanced. The dataset originally was found to be right-skewed.



	CUSTOMERID	SUCCESS	Last Name	First Name	AGE	GENDER	SEATCLASS	GUESTS	FARE
0	1	0	Braund	Mr. Owen Harris	22.000000	M	3	1	7.2500
1	2	1	Cumings	Mrs. John Bradley (Florence Briggs Thayer)	38.000000	F	1	1	71.2833
2	3	1	Heikkinen	Miss. Laina	26.000000	F	3	0	7.9250
3	4	1	Futrelle	Mrs. Jacques Heath (Lily May Peel)	35.000000	F	1	1	53.1000
4	5	0	Allen	Mr. William Henry	35.000000	M	3	0	8.0500
...	...	...	...	...	...	...	...	...	...
886	887	0	Montvila	Rev. Juozas	27.000000	M	2	0	13.0000
887	888	1	Graham	Miss. Margaret Edith	19.000000	F	1	0	30.0000
888	889	0	Johnston	Miss. Catherine Helen "Carrie"	29.699118	F	3	1	23.4500
889	890	1	Behr	Mr. Karl Howell	26.000000	M	1	0	30.0000
890	891	0	Dooley	Mr. Patrick	32.000000	M	3	0	7.7500

Fig Dataset after missing value imputation

III. METADATA EXPLORATION

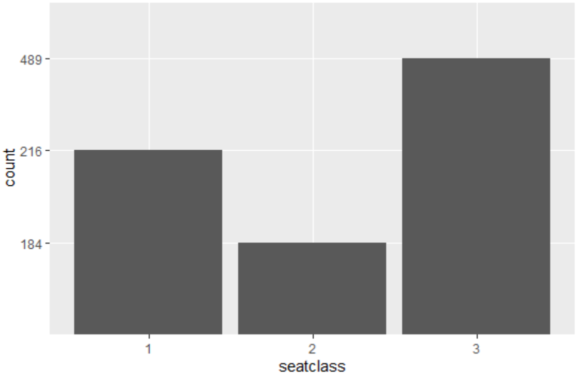


Fig 8 Bar plot for SEAT CLASS

The above bar chart was plot using ggplot2 library from R Studio. It shows that maximum number of passengers prefer to travel by 3rd class. The next most popular class is 1st and 2nd is the least preferred travel class by passengers.

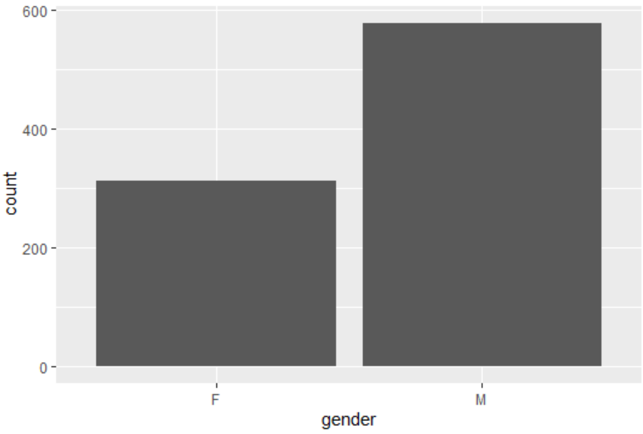


Fig 9 Customer distribution based on Gender

Fig 9 shows that clearly there were more male passengers flying with Airline A than female passengers. The analysis shows that out of total 891 passengers, only 313 were female passengers and rest were male. There are guests as well travelling with the customers, but we don't have their data as of now.

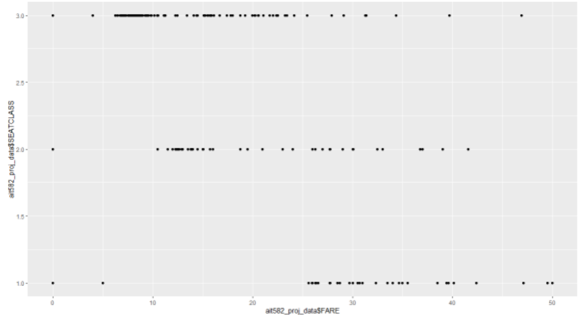


Fig 10 Scatter plot for FARE and SEATCLASS

The above scatterplot shows that SEATCLASS 3 has the minimum fare and mostly passengers have opted for that. Fare for SEATCLASS 2 is not very high neither it is very low. The fare for SEATCLASS 1 is the highest and not many people has opted for it.

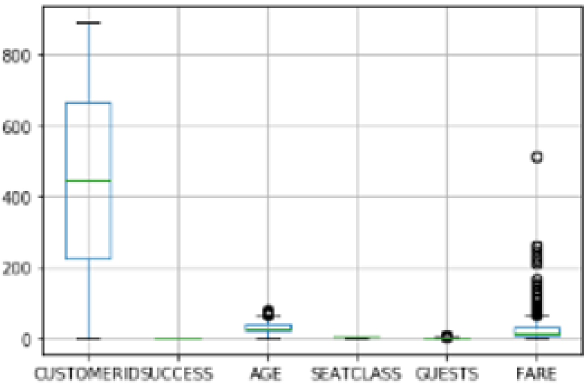


Fig 11 Box Plot for data distribution

The above box plot shows data distribution for variables CUSTOMERID, AGE, SEATCLASS, GUESTS and FARE. As we can see in the plot that variable FARE has a outlier value and a fairly distributed data. Variable AGE does not have a large data distribution. Its values are mainly concentrated. SEATCLASS is a nominal variable so we are not concerned about it in terms of data distribution.

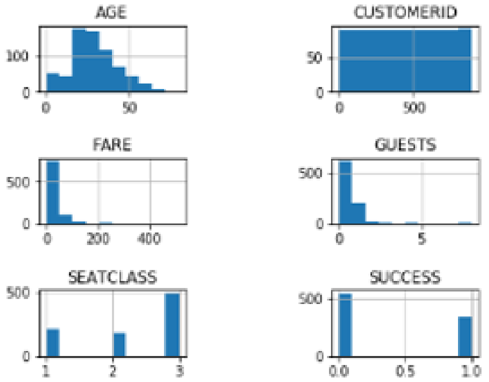


Fig 12 Histogram to check for data skewness

The above histogram is part of the univariate analysis and depicts that the data is largely skewed and required standardization before we can apply any supervised learning model. AGE variable is right skewed and even FARE and GUESTS variable are also right skewed. We cannot get reliable results unless we standardize the data first.

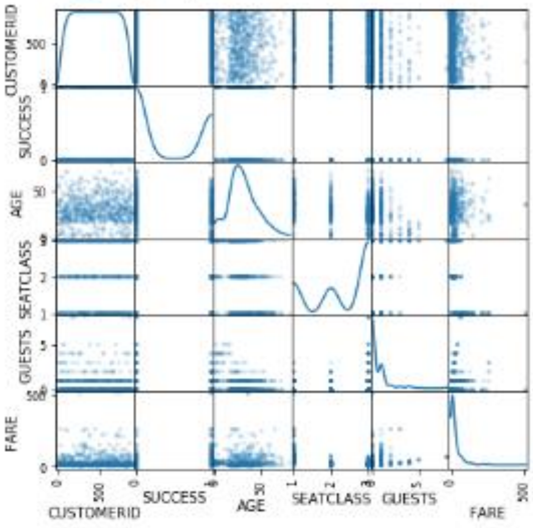


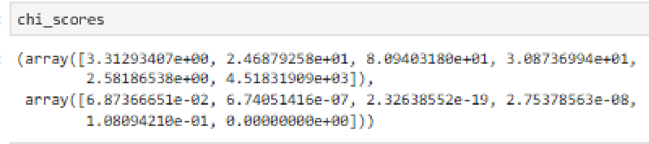
Fig 13 Scatter Plot Matrix

The above scatter plot matrix helps us to identify potential correlation between the different variables. As we can notice from the graph that there is some relation between customer age and fare prices. It looks like lot of customers tend to purchase the lower price tickets and as the person's age is more, they usually buy more expensive tickets or fly in a higher class in flights.

IV. FEATURE SELECTION USING CHI-SQUARE AND WEKA

We need to determine the relationship between the independent category features (predictor) and dependent category feature(SUCCESS). In feature selection, we aim to select the features which are highly dependent on the response. When two features are independent, the observed count is close to the expected count, thus we will have smaller Chi-Square value. So high Chi-Square value indicates that the hypothesis of independence is incorrect. In simple words, higher the Chi-Square value the feature is more dependent on the response and it can be selected for model training. We applied **Label Encoder** to encode 'GENDER'.

Below values were observed after applying Chi-Square test.



```
chi_scores
```

```
(array([3.31293407e+00, 2.46879258e+01, 8.09403180e+01, 3.08736994e+01,
        2.58186538e+00, 4.51831909e+03]),
 array([6.87366651e-02, 6.74051416e-07, 2.32638552e-19, 2.75378563e-08,
        1.08094210e-01, 0.00000000e+00]))
```

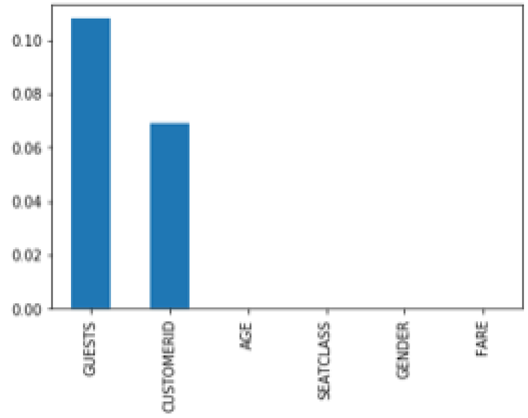
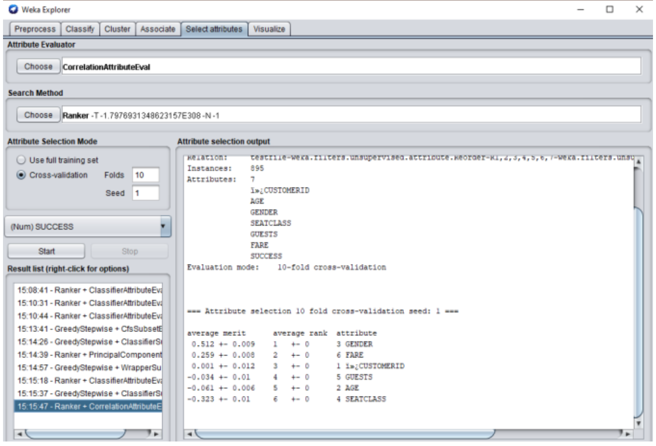


Fig 14 Predictors ranked based on Chi-Square p values



Attribute Evaluator

Choose CorrelationAttributeEval

Search Method

Choose Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

○ Use full training set
● Cross-validation Folds 10
Seed 1

(Num) SUCCESS

Result list (right-click for options)

```
15:08:41 - Ranker + ClassifierAttributeEv
15:10:31 - Ranker + ClassifierAttributeEv
15:10:44 - Ranker + ClassifierAttributeEv
15:13:41 - GreedyStepwise + CfsSubsetE
15:14:26 - GreedyStepwise + ClassifierS
15:14:39 - Ranker + PrincipalComponent
15:14:57 - GreedyStepwise + WrapperSu
15:15:18 - Ranker + ClassifierAttributeEv
15:15:37 - GreedyStepwise + ClassifierS
15:15:47 - Ranker + CorrelationAttributeE
```

Attribute selection output

```
Relation:     testfile-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,7-weka.filters.unsu
Instances:    895
Attributes:   7
              ï»¿CUSTOMERID
              AGE
              GENDER
              SEATCLASS
              GUESTS
              FARE
              SUCCESS
Evaluation mode:    10-fold cross-validation



=== Attribute selection 10 fold cross-validation seed: 1 ===

average merit      average rank  attribute
 0.512 +- 0.009    1   +- 0      3 GENDER
 0.259 +- 0.008    2   +- 0      6 FARE
 0.001 +- 0.012    3   +- 0      1 ï»¿CUSTOMERID
-0.034 +- 0.01     4   +- 0      5 GUESTS
-0.061 +- 0.006    5   +- 0      2 AGE
-0.323 +- 0.01     6   +- 0      4 SEATCLASS
```

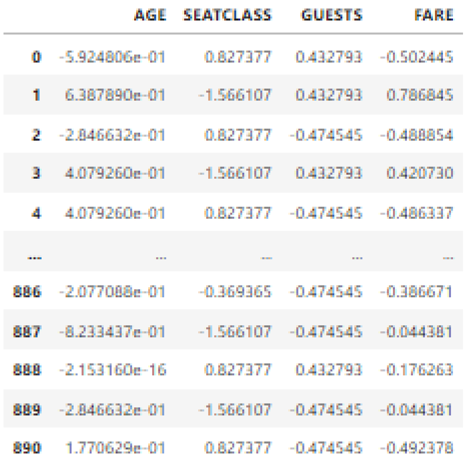
Fig 15 Attribute selection using Weka

CorrelationAttributeEval method evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class.

The top two features according to Chi-Square test and **CorrelationAttributeEval** with 10-fold cross validation are **'GENDER' and 'FARE'**.

V. Data Standardization

Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data. PCA is effected by scale so you need to scale the features in your data before applying PCA. **StandardScaler** was used to standardize the dataset's features onto unit scale (mean = 0 and variance = 1) which is a requirement for the optimal performance of many machine learning algorithms.



	AGE	SEATCLASS	GUESTS	FARE
0	-5.924806e-01	0.827377	0.432793	-0.502445
1	6.387890e-01	-1.566107	0.432793	0.786845
2	-2.846632e-01	0.827377	-0.474545	-0.488854
3	4.079260e-01	-1.566107	0.432793	0.420730
4	4.079260e-01	0.827377	-0.474545	-0.486337
...	...	...	...	...
886	-2.077088e-01	-0.369365	-0.474545	-0.386671
887	-8.233437e-01	-1.566107	-0.474545	-0.044381
888	-2.153160e-16	0.827377	0.432793	-0.176263
889	-2.846632e-01	-1.566107	-0.474545	-0.044381
890	1.770629e-01	0.827377	-0.474545	-0.492378

Fig 16 Dataset after Standardization

Principal Component Analysis

The original data has 4 columns (AGE,SEATCLASS,GUESTS,FARE). In this section, the code projects the original data which is 4 dimensional into 2 dimensions. I should note that after dimensionality reduction, there usually isn't a meaning assigned to each principal component. The new components are just the two main dimensions of variation.

Explained Variance The explained variance tells us how much information (variance) can be attributed to each of the principal components. This is important as while we can convert 4-dimensional space to *2-dimensional* space, you lose some of the variance (information) when we do this. By using the attribute **explained_variance_ratio_**, we see that the first principal component contains 42.2% of the variance and the second principal component contains 30.9% of the variance. Together, the two components contain 73.1% of the information.

VI. ML Algorithm 1 – Decision Trees

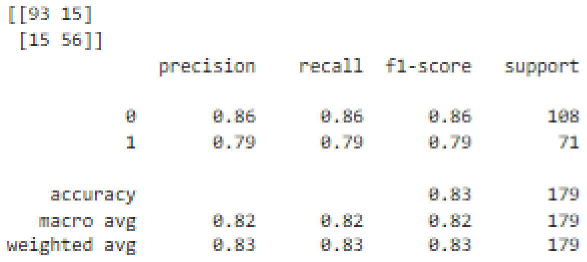
Our goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. After standardizing our dataset, we will now split it into training and test set to train and test our model. The final preprocessing step is to divide our data into training and test sets. The model_selection library of Scikit-Learn contains train_test_split method, which we'll use to randomly split the data into training and testing sets.

Since we are going to perform a classification task here, we will use the DecisionTreeClassifier class.

The fit method of this class is called to train the algorithm on the training data, which is passed as parameter to the fit method. Now that our classifier has been trained, let's make predictions on the test data. To make predictions, the predict method of the DecisionTreeClassifier class is used. At this point we have trained our algorithm and made some predictions. Now we'll see how accurate our algorithm is.

For classification tasks some commonly used metrics are confusion matrix, precision, recall, and F1 score.

Uisng the scikit Learn's classification_report and confusion_matrix methods we evaluate below results:



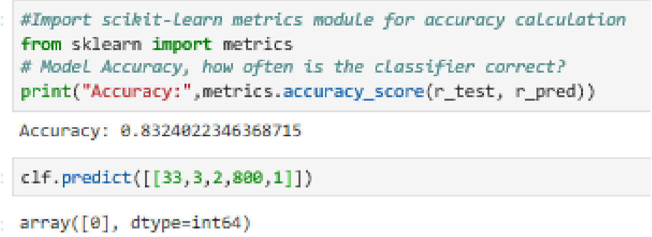
```
[[93 15]
 [15 56]]
              precision    recall  f1-score   support

           0       0.86      0.86      0.86       108
           1       0.79      0.79      0.79        71

    accuracy                           0.83       179
   macro avg       0.82      0.82      0.82       179
weighted avg       0.83      0.83      0.83       179
```

From the confusion matrix, we can see that out of 179 test instances, our algorithm misclassified only 30. This is 83 % accuracy.

VII. ML Algorithm 2 – Random Forests

Random forests as a supervised learning algorithm can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

Random forests is slow in generating predictions because it has multiple decision trees. Whenever it makes a prediction, all the trees in the forest have to make a prediction for the same given input and then perform voting on it. This whole process is time-consuming. So, After splitting our dataset, we will train the model on the training set and perform predictions on the test set. After training, we check the accuracy using actual and predicted values. We received an accuracy of 83.2%



```
#Import scikit-learn metrics module for accuracy calculation
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(r_test, r_pred))
```

```
Accuracy: 0.8324022346368715
```

```
clf.predict([[33,3,2,800,1]])
```

```
array([0], dtype=int64)
```

ROC Curve

A ROC curve plots the performance of a binary classifier under various threshold settings; this is measured by true positive rate and false positive rate. If the classifier predicts "true" more often, it will have more true positives (good) but also more false positives (bad). If the classifier is more conservative, predicting "true" less often, it will have fewer false positives but fewer true positives as well. The ROC curve is a graphical representation of this tradeoff.

A perfect classifier has a 100% true positive rate and 0% false positive rate, so its ROC curve passes through the upper left corner of the square. A completely random classifier (ie: predicting "true" with probability p and "false" with probability 1-p for all inputs) will by random chance correctly classify proportion p of the actual true values and incorrectly classify proportion p of the false values, so its true and false positive rates are both p. Therefore, a completely random classifier's ROC curve is a straight line through the diagonal of the plot.

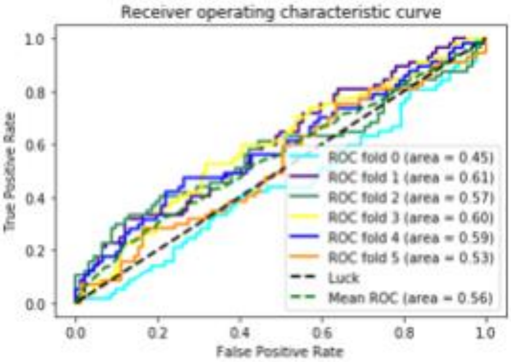


Fig 17 ROC Curve with cross validation

The AUC (Area Under Curve) is the area enclosed by the ROC curve. A perfect classifier has AUC = 1 and a completely random classifier has AUC = 0.5. Usually, our model will score somewhere in between. The range of possible AUC values is [0, 1]. However, if the AUC is below 0.5, that means we can invert all the outputs of the classifier and get a better score, so we did something wrong. The mean ROC was found to be 0.56 which is only slightly better.

VIII. CONCLUSION

Airline dataset was analyzed, and descriptive metadata was extracted which provided us information about Gender and Age of the customers. The missing values in the Age column were imputed using mean value imputation. Top attributes were identified after imputing missing values which were 'GENDER' and 'FARE'. Principal component analysis was performed for dimensionality reduction and visualization of the attributes in a 2-D space. Top features were selected. Then the ML algorithms – Decision Trees and Random Forests were created and tested for accuracy. Finally, the ROC and AUC curve was plotted and analyzed to validate model accuracy. It was found to be around 0.56.

REFERENCES

[1] Rafatirad, S.(2020). 4.1 - Metadata Extraction Techniques. Retrieved from https://mymasonportal.gmu.edu/webapps/blackboard/execute/displayLearningUnit?course_id=_381834_1&content_id=_10220365_1

[2] Rafatirad, S(2020). Data Distributions and Visualizations. Retrieved from https://mymasonportal.gmu.edu/webapps/blackboard/execute/displayLearningUnit?course_id=_381834_1&content_id=_10220524_1

[3] Fawcett. T. (March 16,2004). ROC Graphs: Notes and Practical ConsiderationsforResearchers. Retrieved from http://www.blogspot.udec.ugto.saedsayad.com/docs/ROC101.pdf