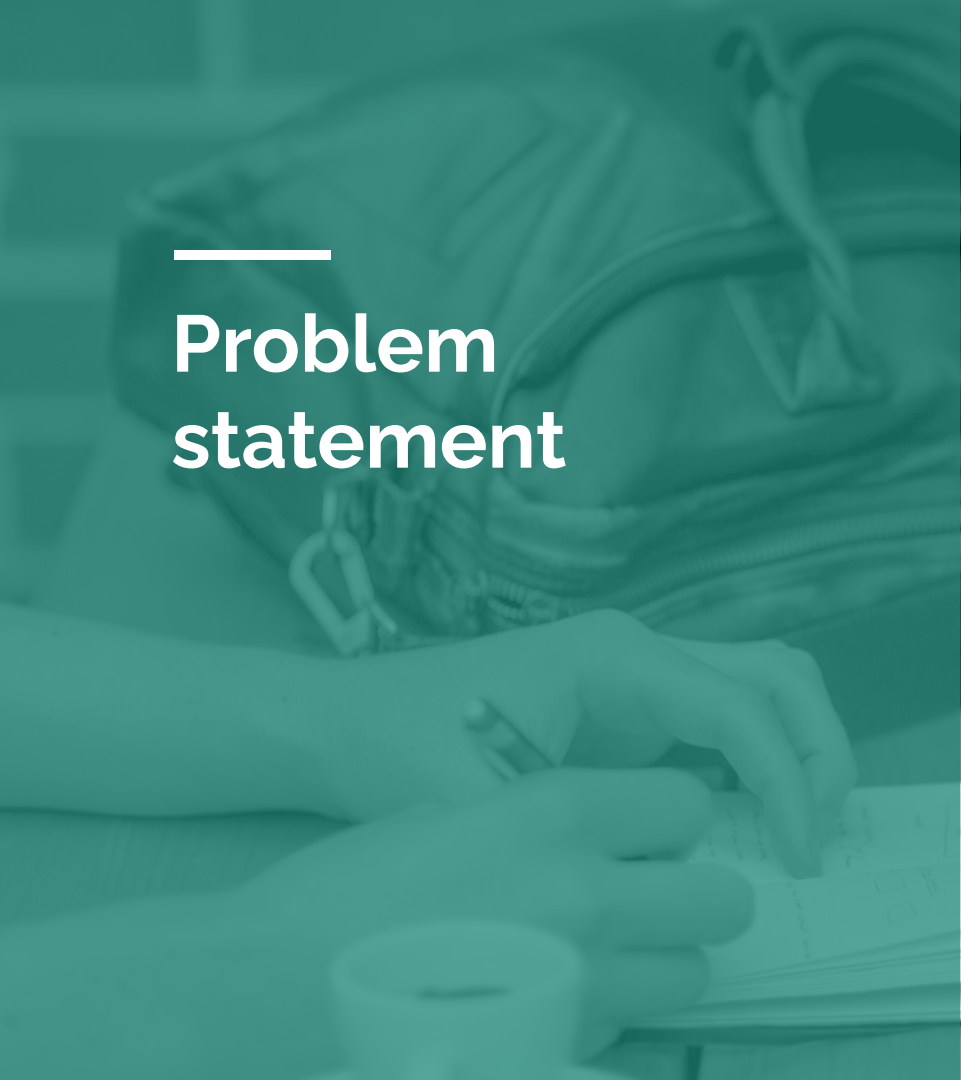
Data Science WorkFlow

Problem statement, Analysis, Approach
and Results.

Outline

The Problem



Problem statement

Using airline passenger satisfaction survey dataset to predict the customer satisfaction with a binary classification model.

The data concerns customer satisfaction, to be predicted in terms of 5 multivalued discrete and 20 continuous attributes.

Prediction Target: satisfaction

Description of DataSet

Attribute Information:

- Gender: Gender of the passengers (Female, Male).
- Customer Type: The customer type (Loyal customer, disloyal customer).
- Age: The actual age of the passengers.
- Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel).
- Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus).
- Flight distance: The flight distance of this journey.
- Inflight wifi service: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5).
- Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient.
- Ease of Online booking: Satisfaction level of online booking.
- Gate location: Satisfaction level of Gate location.

contd...

Description of DataSet

Attribute Information:

- Food and drink: Satisfaction level of Food and drink.
- Online boarding: Satisfaction level of online boarding.
- Seat comfort: Satisfaction level of Seat comfort.
- Inflight entertainment: Satisfaction level of inflight entertainment.
- On-board service: Satisfaction level of On-board service.
- Leg room service: Satisfaction level of Leg room service.
- Baggage handling: Satisfaction level of baggage handling.
- Check-in service: Satisfaction level of Check-in service.
- Inflight service: Satisfaction level of inflight service.
- Cleanliness: Satisfaction level of Cleanliness.
- Departure Delay in Minutes: Minutes delayed when departure.
- Arrival Delay in Minutes: Minutes delayed when Arrival.

Training Size: 103904 Examples, Testing Size: 25976 Examples

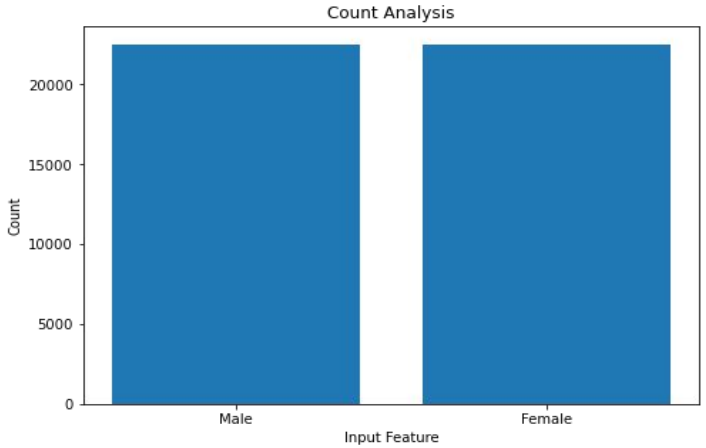
Exploratory Analysis

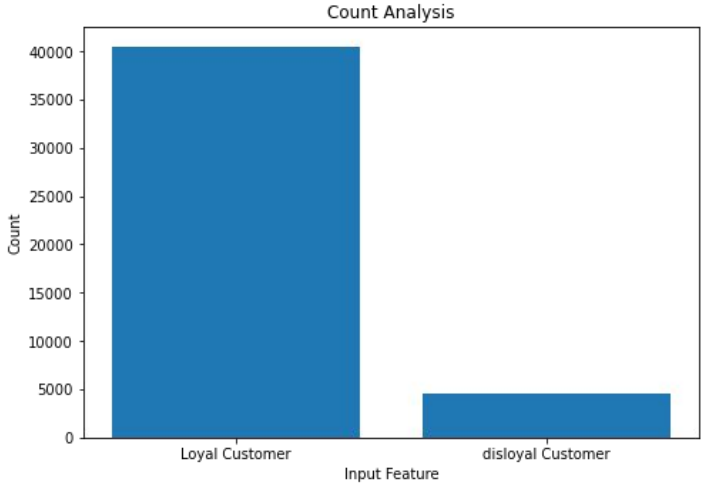
Exploratory Analysis

01

Analyzing the categorical discrete columns, for the satisfaction index.

- Male and Female are almost equal, so gender is not a factor
- Loyal Customers are always more satisfied than disloyal customers



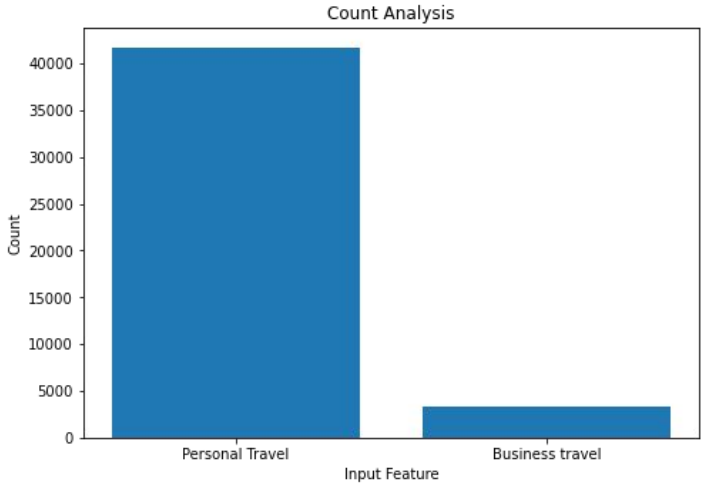


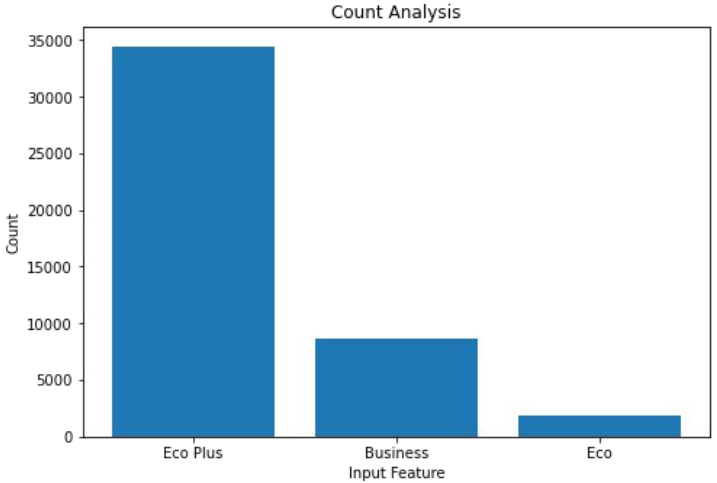
Exploratory Analysis

01

Analyzing the categorical discrete columns, for the satisfaction index.

- Customer travelling for personal reasons end up more satisfied than the ones travelling for business.
- Eco Plus customers are way more often satisfied than the business class, whereas Eco - passengers are the least satisfied.





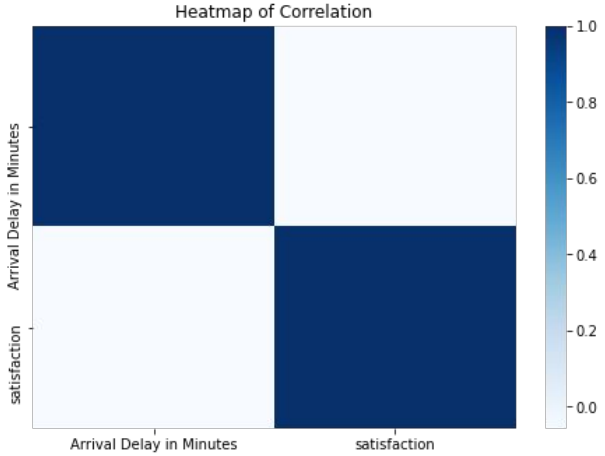
Exploratory Analysis

02

Since the only column with null values is 'Arrival Delay in Minutes', we'll check its correlation with 'satisfaction'

Correlation:

Input Feature Arrival Delay in Minutes has a low correlation with satisfaction, and hence can be imputed to replace null values.

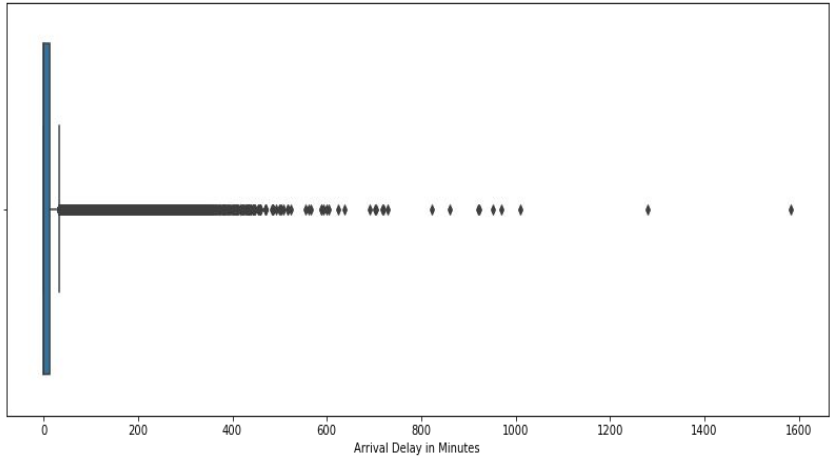


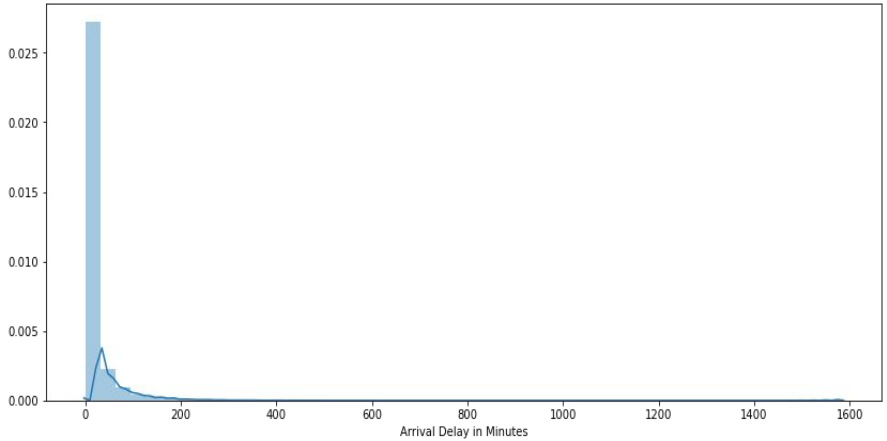
Exploratory Analysis

03

Before performing imputation, we'll have to check the skew of the dataset.

Our dataset is right skewed, has multiple outliers towards the right, as evident from the plots

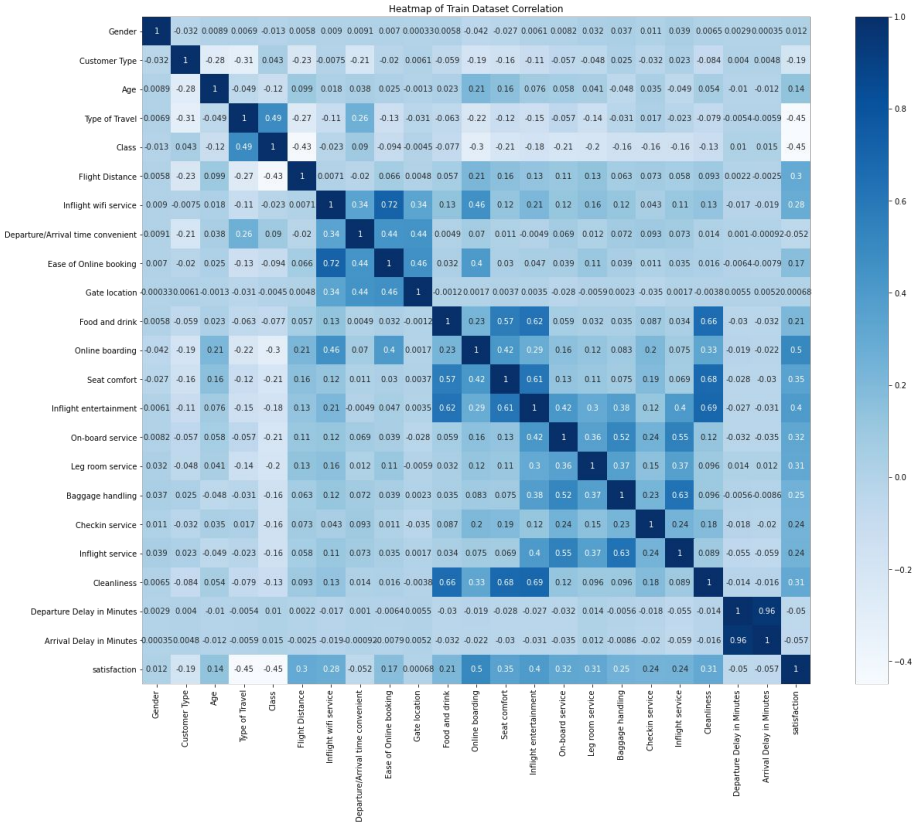




Exploratory Analysis

04

We can also check the correlation heatmap to figure out the correlation between the input features and the satisfaction, to get an idea of the more influential features.



Please check the notebook for a clearer heatmap.

Approach

Data Preprocessing

- From our analysis, it is evident that although gender is not an influential feature, still, features like 'Class', 'Customer Type', and 'Type of Travel' are of necessity, hence we'll have to encode the categorical columns (including 'satisfaction') to numerical variables using **LabelEncoder()**.
- We have also found that the correlation between 'satisfaction' and 'Arrival Delay in Minutes' to be less, therefore we can Impute to fill in the null values using **SimpleImputer()**. We'll be using the **median()** strategy as we've found the data to be right skewed with multiple outliers on the right side, which can cause an adverse effect on mean().
- Lastly we've found from the heat map, that columns -- 'Gate location', 'Gender', and 'Departure/Arrival Time Convenient'. Therefore, we'll be dropping these columns as they won't be of much use to our prediction model. However 'Departure/Arrival Time Convenient', is a rating column, and hence we'll be keeping the same.
- Finally, after data has been preprocessed, we've normalized the data to 1NF (0-mean; 1-standard deviation) using **StandardScaler()**.

Description

Now that the data has been preprocessed, we'll be creating 4 different models to fit and train our data, so that we can predict the satisfaction (0- neutral/dissatisfied, 1-satisfied)

Models:

- Random Forest Classifier
- XGBoost
- Deep Neural Network with 2 hidden layers and 1 output layer
- Deep Neural Network with 4 hidden and 1 output layer with BatchNormalization() layer to bring data to 1 NF.

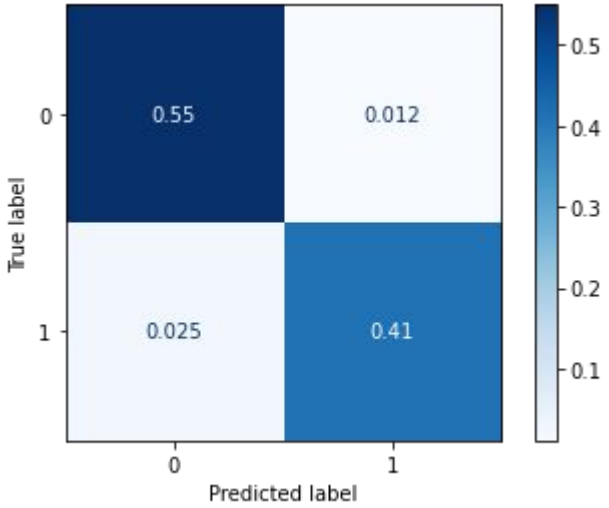
Approach

01

We'll use the **RandomForestClassifier**() model first, with no parameter.

We've calculated the error as ROC, r2_score, and accuracy.

The model gives us a 96.3% accuracy with an r2_score and ROC_AUC score of 0.85 and 0.96 respectively.

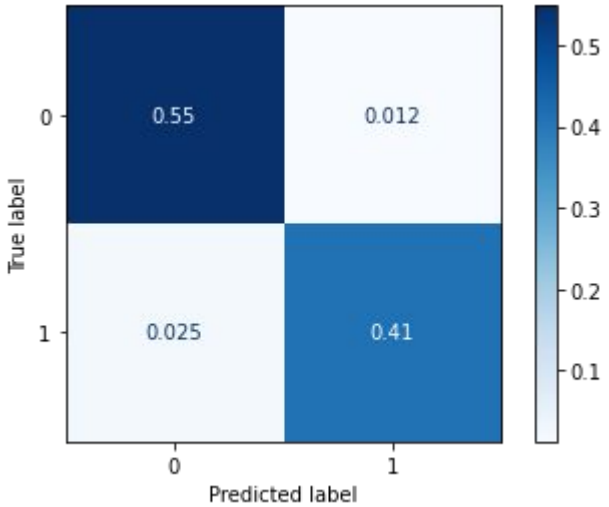


Approach
02

We'll use the **XGBoost**() model first, with no parameter.

We've calculated the error as ROC, r2_score, and accuracy.

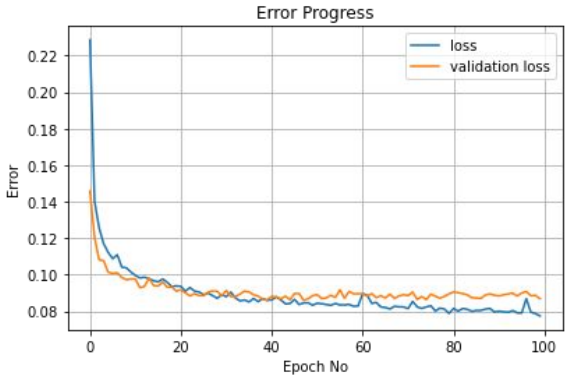
The model gives us a 96.2% accuracy with an r2_score and ROC_AUC score of 0.848 and 0.96 respectively.



Approach 03

Deep Neural Network with 2 hidden layers and 1 output layer using Keras API, **Sequential().**

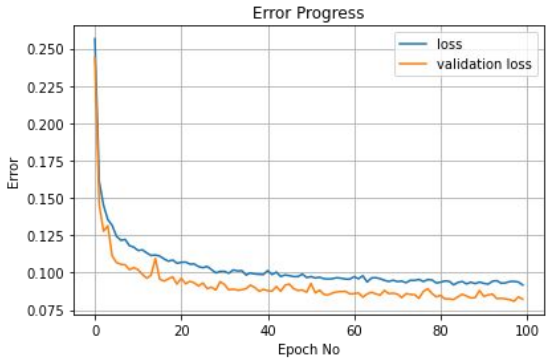
We've used the accuracy metric. The model gives us a 96.36% accuracy with a steady loss decline and a final loss of 8%.

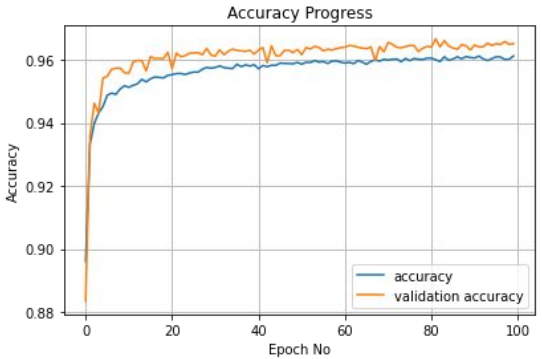


Approach 03

Deep Neural Network with 4 hidden layers and 1 output layer, and **BatchNormalization()** using Keras API, **Sequential().**

We've used the accuracy metric. The model gives us a 96.28% accuracy with a steady loss decline and a final loss of 8%.





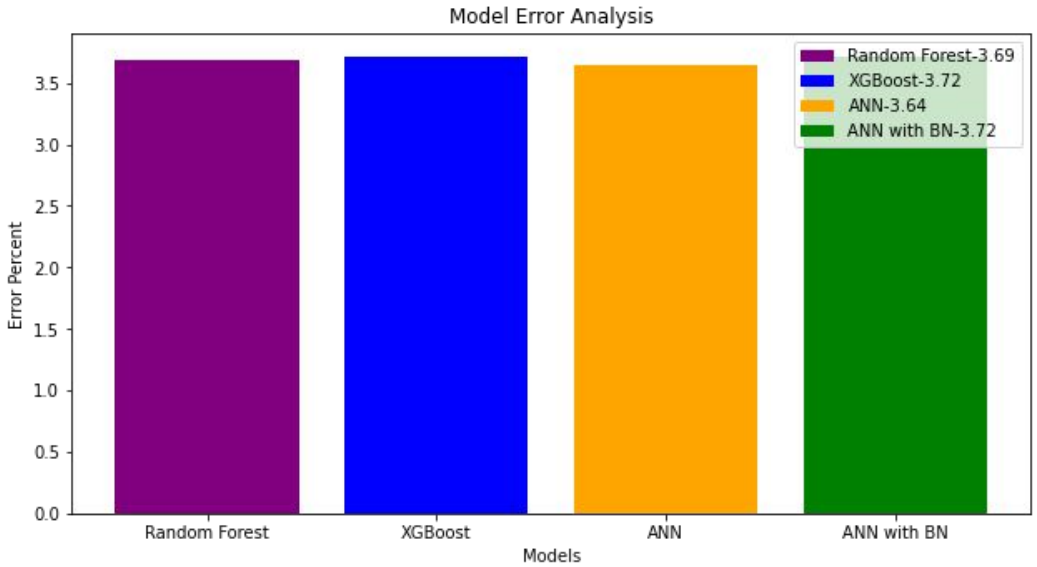
Results

Conclusion

To conclude, all 4 of our models have given us similar results.

- Random forest model: 96.31%
- XGBoost model: 96.28%
- DNN (2 layer): 96.36%,
- DNN (4 layer): 96.28%

However we can see that the 2 layered DNN model has the highest accuracy, marginally.



Thank You.

References

Please check the jupyter notebook: Customer Satisfaction.ipynb

Dataset: https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction/