Airlines Customer Satisfaction (Classification)

Rutvik Patel

May 2, 2023

Contents

List of Figures

1 Introduction

Airlines play a vital role in the transportation industry, connecting people and businesses across the globe. As a result, the satisfaction of airline customers is crucial to the success of airline companies. In today's competitive world, airline companies need to understand what factors influence customer satisfaction, and how they can improve their services to meet their customer's needs. This data set will help to identify key factors behind customer's satisfaction in which airline can improve. This problem comes under classification problem. Here, we will train several models on this data set and try to find few key factor of it.

2 Dataset

The **"Airlines Customer Satisfaction (Classification)"** was obtained from the Kaggle Data Science. The size of our data set is**"25976 X 25"** in which we have 25976 rows, 24 features columns and 1 result(label) column. our data set contain variety of factors like, **'Gender', 'Customer Type', 'Age', 'Type of Travel', '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', 'On-board service', 'Leg room service', 'Baggage handling', 'Checkin service', 'Inflight service', 'Cleanliness', 'Departure Delay in Minutes', 'Arrival Delay in Minutes'**. Mostly, all attributes are having input value between 0 to 5 and rest of attribute like 'age','gender', 'Type of Travel', 'Class' are having different string values.

2.1 Visualization of the distribution of each input features

here we are going to visualization row data (without processing) and distribution of out data among two classes.

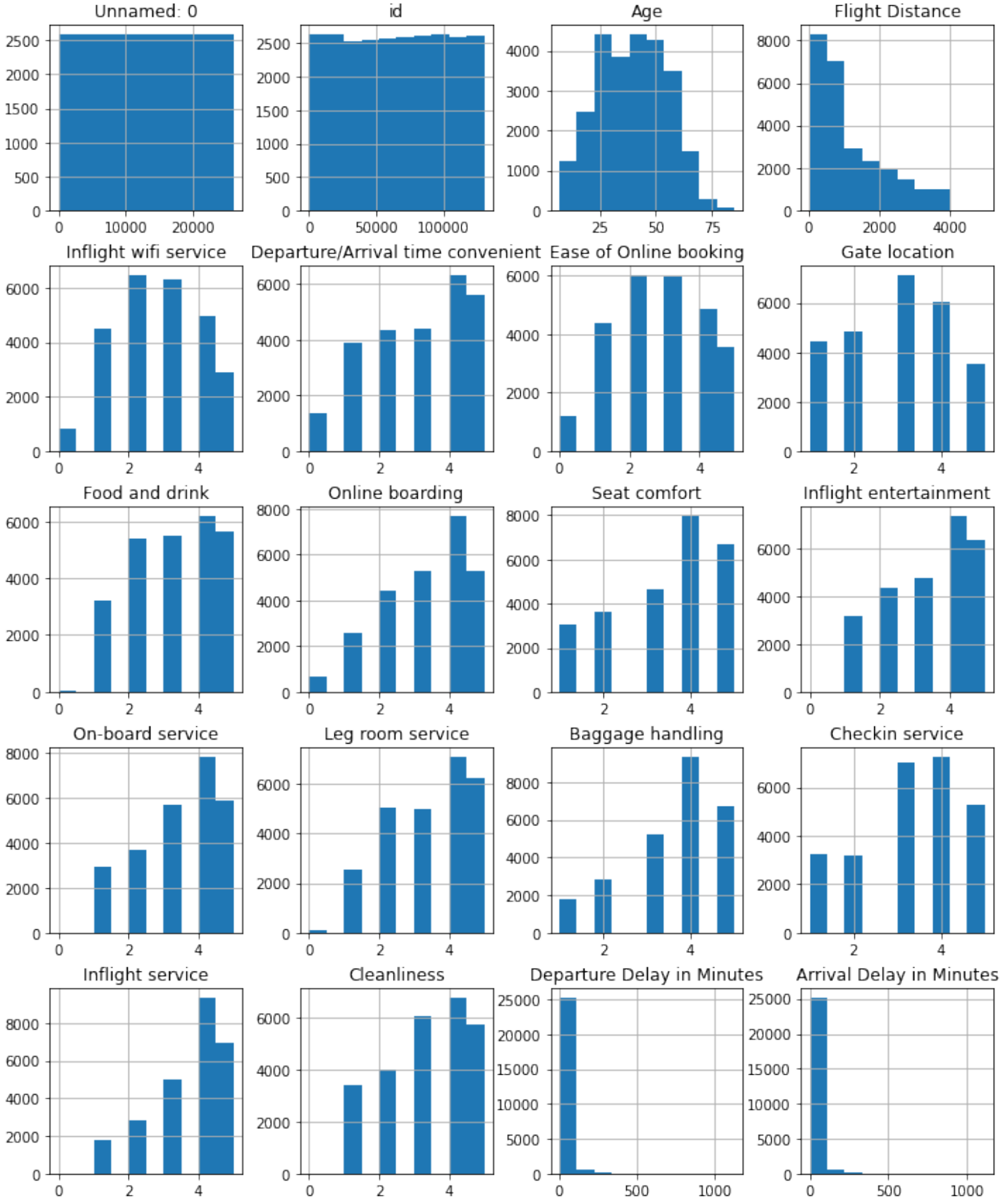

Figure 1: Input Data Distribution - Before Normalization

2.2 Distribution of the output labels

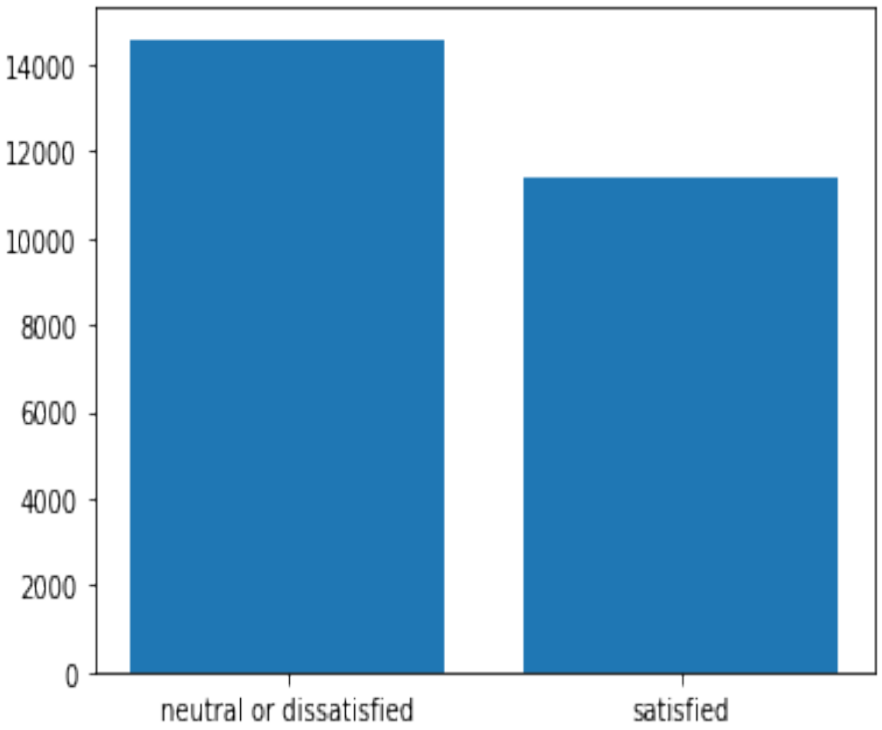

Figure 2: Output Data Distribution

3 Data Pre-Processing

3.1 Finding Null values

In, this phase firstly, we are finding null value(empty cell) in our data set. If null value is present then we will fill it with the use of "mean()" function. Otherwise, with null value we are not able to gain best accuracy of our model.

3.2 Removing unnecessary columns

In, our data set we have some unnecessary columns like "index number" and "customer ID". We will identify those and remove it from our data set in terms of reducing load of our model.

3.3 Data Normalization

Data prepossessing is essential before data mining to address the non-uniform distribution of data. To achieve this, normalization techniques are used to make the optimization problem more numerically stable and improve training. Normalization helps to ensure all values lie between 0 and 1 and outliers are visible within the normalized data. There are two normalization techniques available, each with its own consequences, but either technique can be used for now. bt we are going to use Mean Max Normalization for our project,

Mean Normalization Formula

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

3.4 Visualization of of each input features after normalization

After performing normalization our data is going to look like...

Figure 3: Input Data Distribution - After Normalization

4 Modelling

In this project we are going to use "Tensorflow" to create our neural network model and we will those models on our data set and obtain our goal of this project. At very first we will over-fit our data in different different models. what is over-fitting? and why Over-fit? Over-fitting means without splitting our data in train and test set we will feed whole data to model and try to get 100% accuracy or near by that it help use to define our actual model which we are going to use farther.

4.1 Selected Neural Network Architecture For Over-Fitting

In the beginning state,we are going to start with base line model and then we are increasing layers and epochs until we reach our goal state. the performances of all models are given in table below.

4.1.1 Performance Table

Model	epochs	Loss	Accuracy
1	100	0.4715	0.8516
2-1	500	0.4229	0.8515
18-8-4-1	100	0.1467	0.9414
32-16-8-4-1	100	0.1653	0.9331
5-64-32-8-4-1	250	0.2230	0.9194

Table 1: Performance comparison for different hidden layers

4.2 Learning Curve of Neural Network Architecture

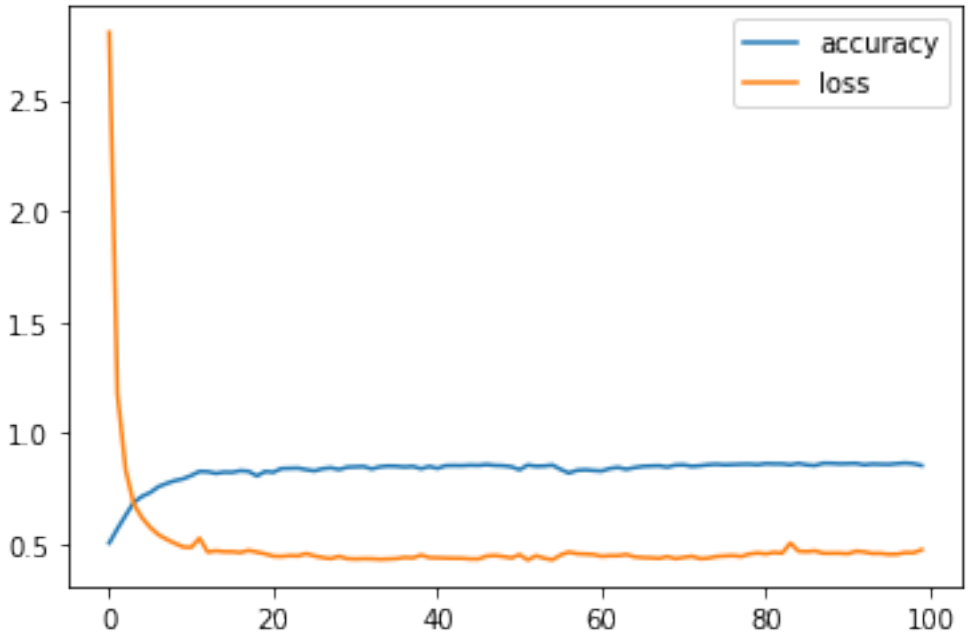

Figure 4: Curve For Model 1

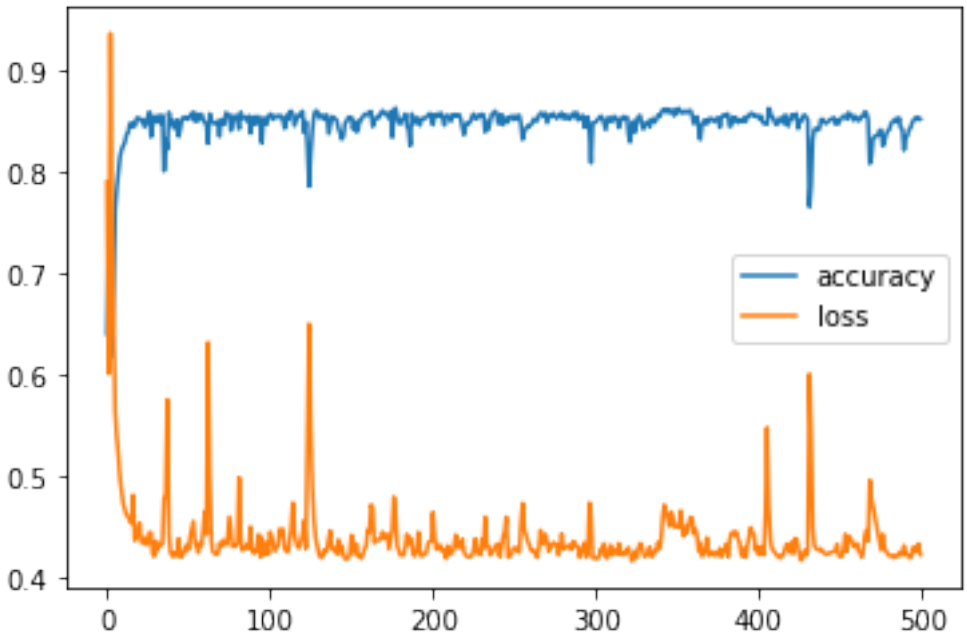

Figure 5: Curve For Model 2-1

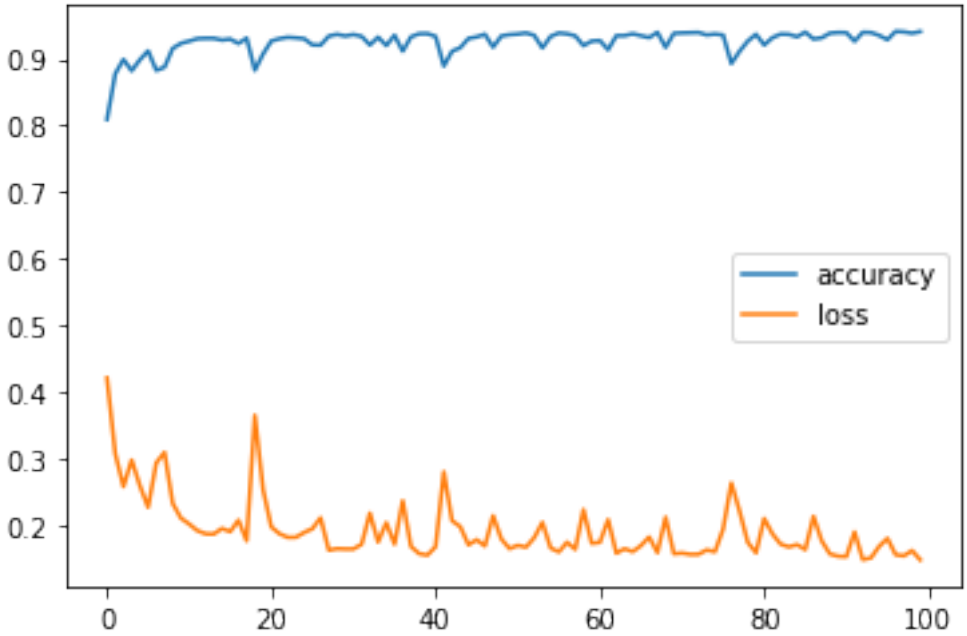

Figure 6: Curve For Model 18-8-4-1

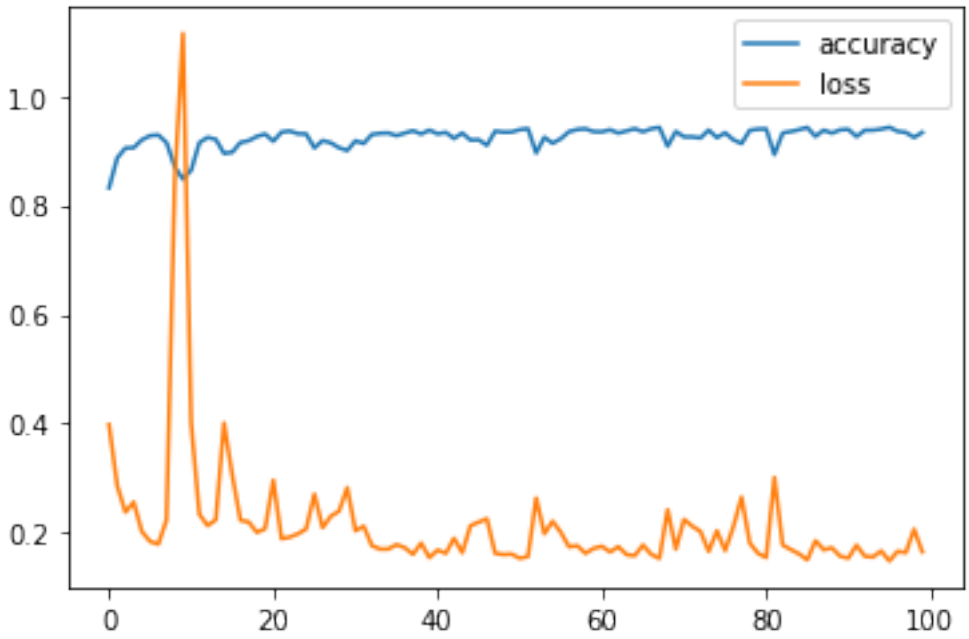

Figure 7: Curve For Model 32-16-8-4-1

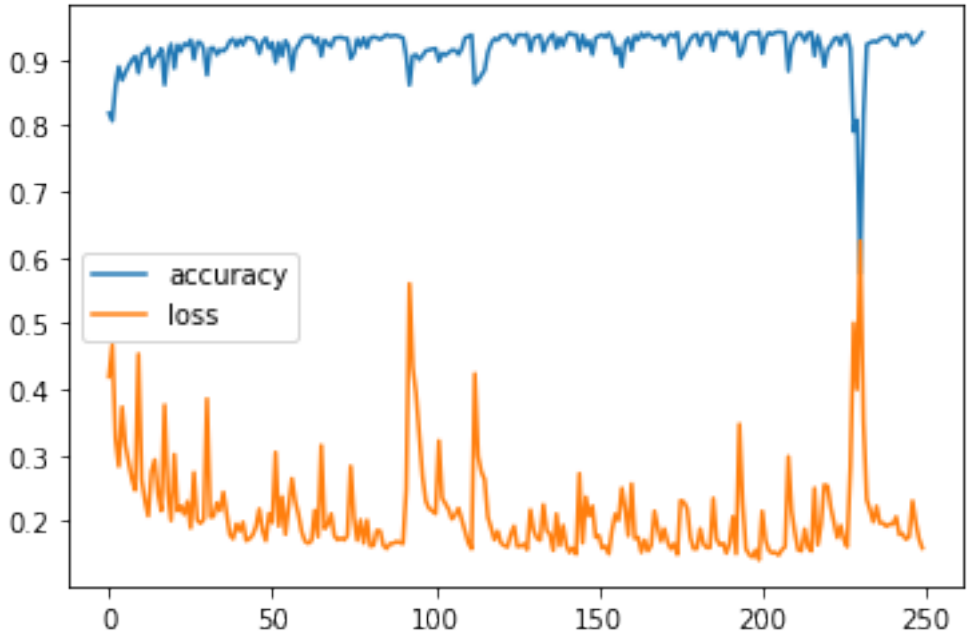

Figure 8: Curve For Model 5-64-32-8-4-1

5 Model Selection

In this phase we will split our data and feed it to different different models to find the best fit model for our project.

5.1 Data Splitting

here, we will take our data set and shuffle it. Because, it may possible that top part of data having same label and rest of having same but by shuffling it we will get almost same amount of data from both the class in both set. After shuffling we will split is in two part 80% of data is going to be training am rest 20% is going to test set. with the use oh training set we will train our model and we will test it with test set.

5.2 Creating Model

Now, we have our both the set, So, we will train and test our model.The performance of all model is given below.

Model	loss	accuracy
2-4-1	0.2261	0.9067
6-8-1	0.1409	0.9394
10-8-1	0.1413	0.9386

5.3 Test Accuracy

A useful tool when predicting the probability of a binary outcome is the Receiver Operating Characteristic curve or ROC curve. The area covered by the curve is the area between the red line and the axis. This area covered is AUC. The bigger the area covered, the better the machine learning models are at distinguishing the given classes. In other words, the AUC can be used as a summary of the model skill. The ideal value for AUC is 1.

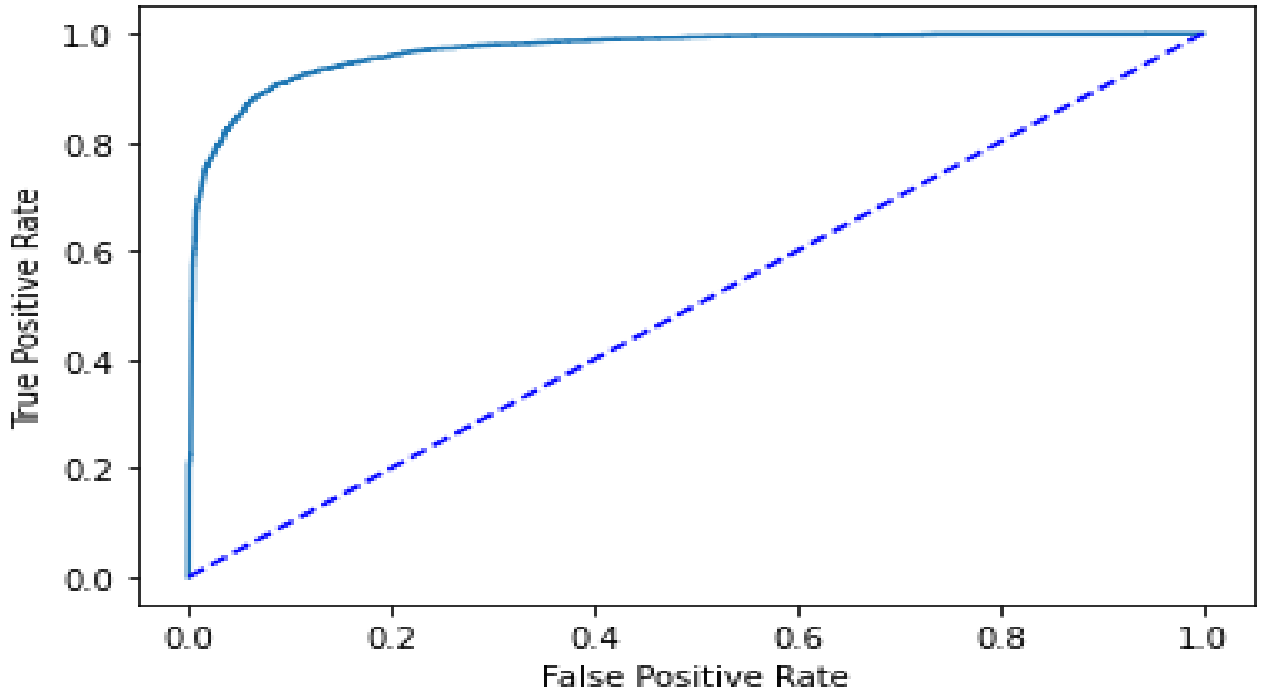

Figure 9: ROC Plotting For Model 2-4-1

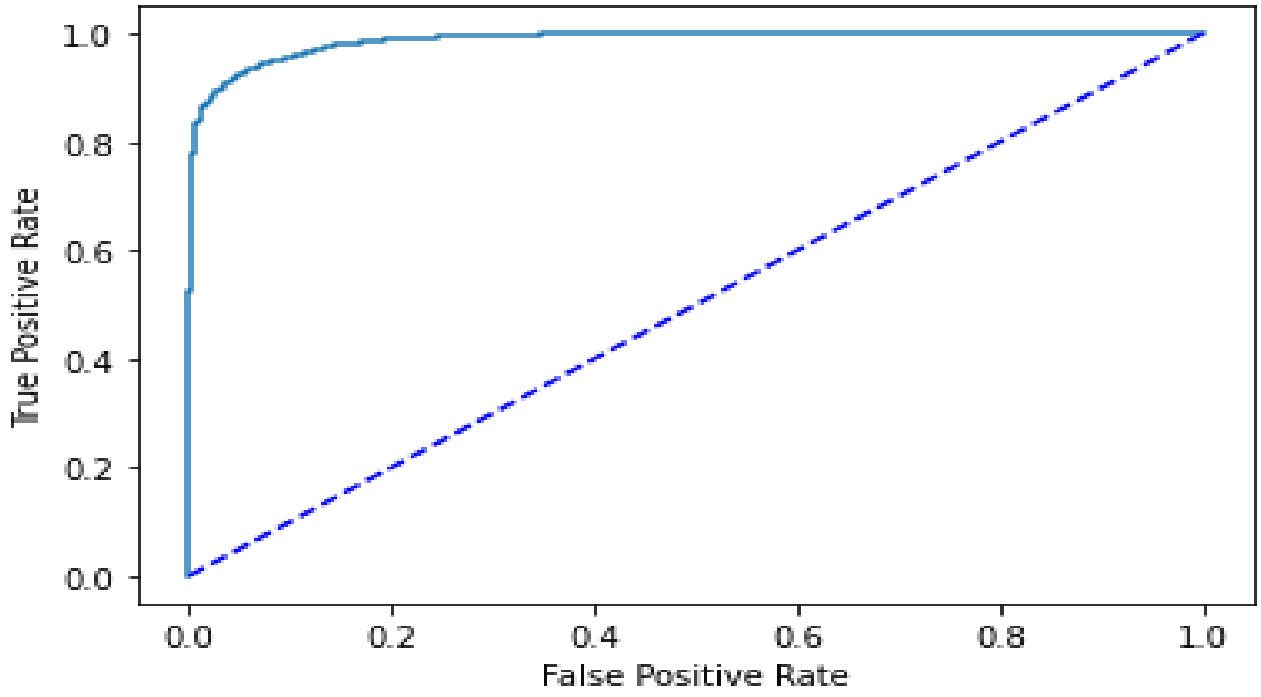

Figure 10: ROC Plotting For Model l 6-8-1

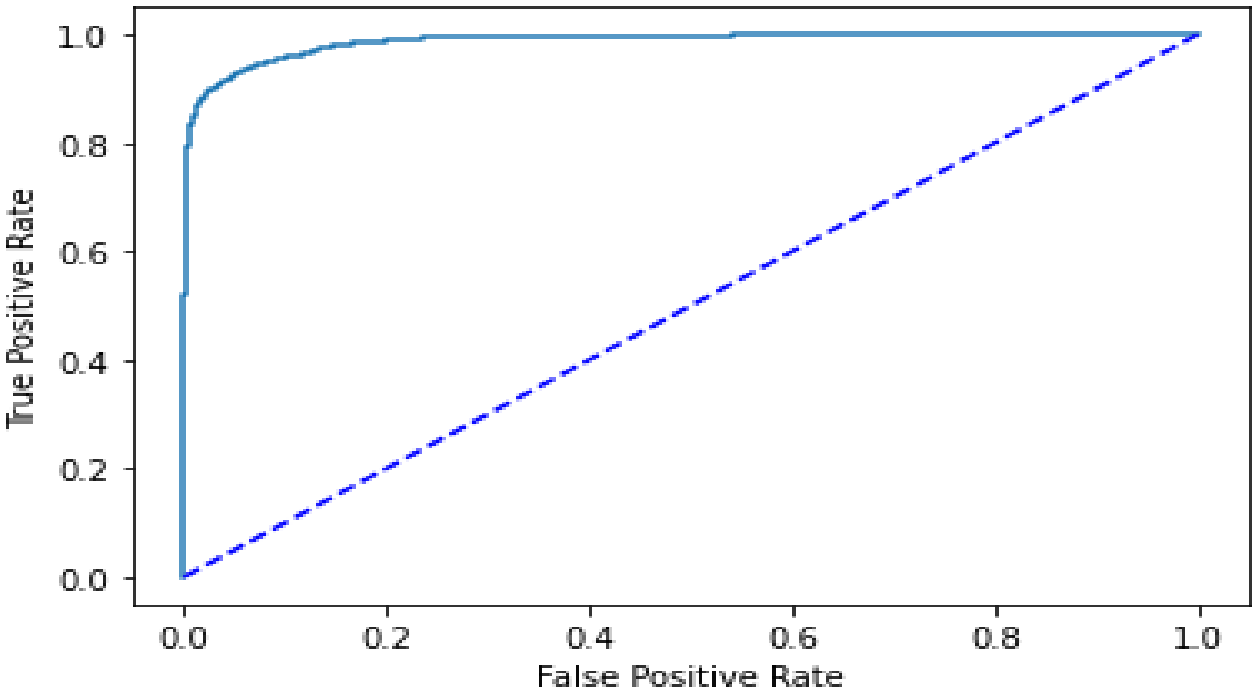

Figure 11: ROC Plotting For Model 10-8-1

6 Feature Importance Analysis

As of now, we tried different models with the number of epochs, and we know which one we have to use in further project. Now we will select one model and try to find importance of each and every features of out set.

6.1 Importance Of Individual Features

We are going to take each individual feature and train our model on that and try to find accuracy. As much high accuracy as high importance. The importance of each attributes are shown in below bar graph.

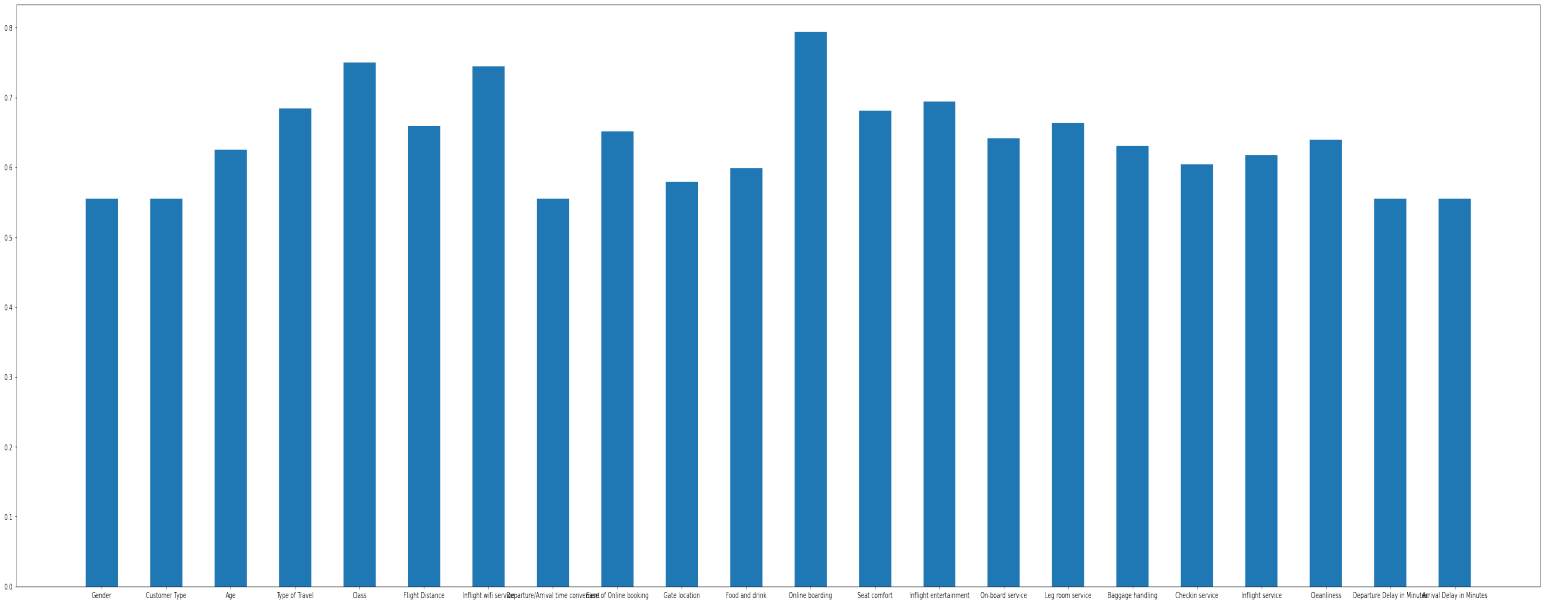

Figure 12: Significance Of Individual Features

6.2 Performance after removing less important features

Now, we are going to remove less importance feature one at the time. after removing feature we will train and test our model and analyze that is there and major drop in accuracy or not? you can find that answer in given graph.

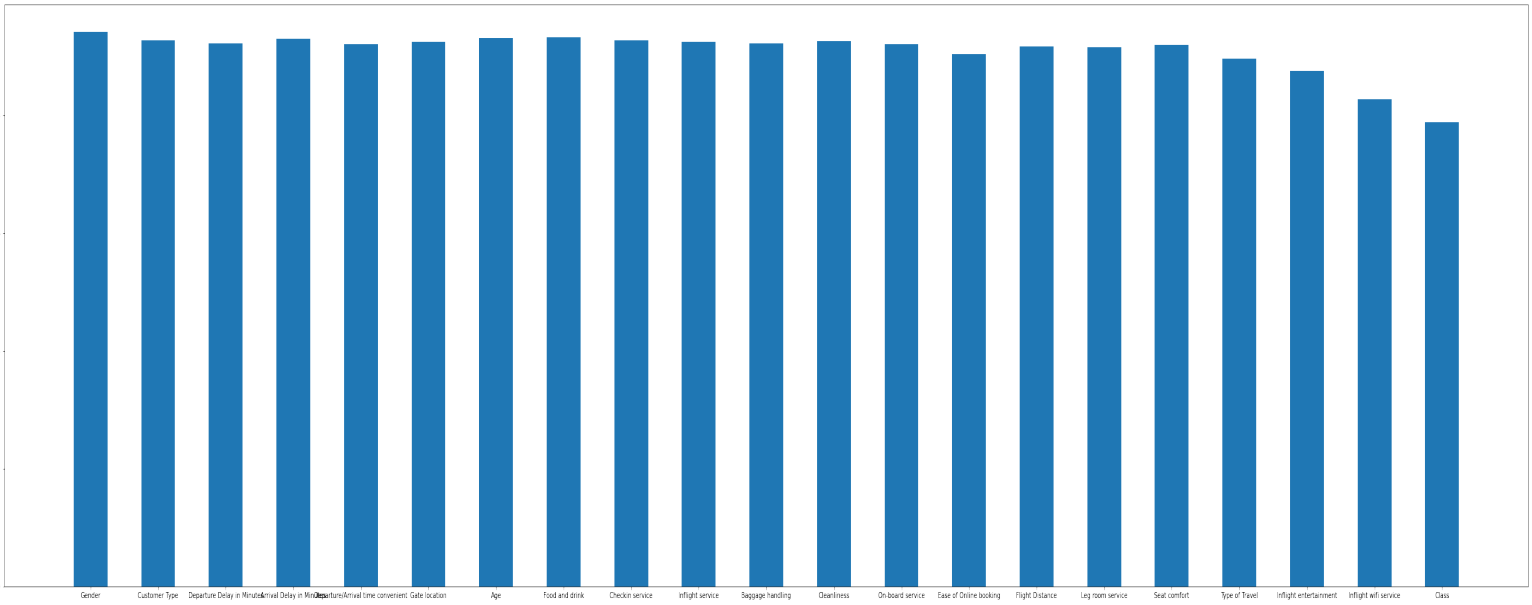

Figure 13: Accuracy After Removing Individual Features

7 Challenges Faced

It is funny but truth, the main challenge I face in entire project is to find good data set I don;t know why but it's true. There are other challenges, one of them was ROC curves but thanks to video link which was given by professors and internet which help me to solve that challenges.

8 Conclusion

To summarize, I can say that this project is very good, effective and fun way to learn and explore neural network models. in this project we started with finding data set to clean it up and make it ready to play with it. In next phase create model to achieve 100% accuracy then try different model and select which is best for you. and at the end play with attributes and find which one is very use full and which one is less use full because "each data is use full".