**Airline Recommendation prediction Based on Customer Review and rating using Ensemble ML and Deep learning models**

**ABSTRACT**

With globalization, the customer's needs and expectations on traveling are varying and the airlines service providers must provide good services, comfort on travelling, food, ground service and many. With the usage of the internet, the customers have freedom to choose the airline based on the other customer reviews and ratings. There are a number of features to be considered to choose the airline, thus there is a need for customers to automate the process of airline recommendation prediction. The recommendation provided by the already travelled customers as 'yes' or 'no' based on the various services offered by the airlines. This work solves the challenges in choosing the appropriate airline for the customers based on the reviews and ratings. The proposed work used an ensemble machine learning algorithm and a deep learning algorithm, Convolutional Neural Network to predict the airline. They are used in different contexts such as rating and reviews. The dataset used is web scrapped from www.airlinequality.com.The reviews based prediction used Natural Language process (NLP) for pre-processing the data. These two context (reviews and ratings) are studied using the ensemble ML model, AdaBoost and CNN model.

**CHAPTER 1**

**INTRODUCTION**

In this internet era, online reviews play major roles in choosing a product or services offered by the companies. The customer chooses a product or services offered based on the ratings or reviews provided to the product or services. Thus understanding the reviews and rating helps in choosing a particular service for the users. Service providers analyzing these data to improve the client's needs and expectations and resolve them will enhance the business.

Online comments are more powerful these days with emerging online shopping by users around the world, the comments are most influencing factors to make decision for users to with the services or products. Analyzing the reviews gives companies to know about the customer satisfaction they delivering and take actions to improve it. As the airline business is more competitive, there are many factors the company has to consider, including price of flight tickets, services they provide and many more. The customer service plays major role for choosing particular airline, the users prefer good services from airlines.

Thus airline companies are in position to analyze the customer reviews to ensure the customer satisfaction to grow their business, similarly, users analyze the reviews to choose the best airlines and get their services.

Artificial intelligence plays major role in major industries like health care, finance, agriculture, ecommerce, banking and more. The role of artificial intelligence in airline industry is enormous. The machine learning and deep learning techniques were studied on different contexts like flight delay prediction and flight price prediction and demand prediction etc. There are less work attempted on review and rating based airline prediction.

It is challenging for users to book airline with good services, they opt to see the reviews and ratings given by other users. The review based analysis requires text analysis, thus Natural Language processing is widely used. The rating based analysis required machine learning algorithm to learn and predict the airline as binary classification.

This project is aimed to solve the above issues by analyzing the reviews using Natural Language Processing and Machine learning models to get the airline recommendation predicted as 'Yes' or 'No'.

Airlinequality.com

Web-scrap data for airline

Dataset

Rating based prediction

Review based prediction

Predict Airline

Classification ADB/CNN

Data Pre-processing

Predict Airline

Classification ADB/CNN

**Figure: Overview of Airline prediction**

The above figure represents the overview of proposed project airline prediction. The dataset is web scrapped from www.airlinequality.com. The dataset has rating as well as review for each airline, these dataset pre-preprocessed to apply ML and DL models to predict the airline recommendation as 'Yes' or 'No', where Yes represents that it is recommended for users and No represents not recommended.

**CHAPTER 2**

**LITERATURE SURVEY**

In this chapter, the literature survey on airline price, demand, delay prediction and related work are discussed. Though there are few works available in these topics, airline recommendation prediction is still not explored.

Flight delay prediction using deep learning model Long Short Term Memory (LSTM) is proposed in [1], the work addressed as big aviation data problem to predict the time delay. The dataset is prepared with automatic dependent surveillance broadcast (ADS-B) technique, the flight current state is captured to create the dataset. These data integrated with climatic conditions, airport schedule and flight schedule the delay is predicted. The model proposed were Random forest and LSTM model for delay prediction. Experimental results showed that LSTM model has achieved the highest accuracy of 99% for flight delay prediction.

Airline delays prediction due to weather condition is proposed in [2], the delays of individual flight is predicted using machine learning models, 10 fold cross validation is used to identify the pattern of unseen data. This paper solves the problem of identifying the relation between the normal variables and delay parameters through machine learning models. Experimental results showed that Random forest has achieved the highest accuracy of 83.4%.

Airline delay with cost sensitive approach was proposed in [3], which computes the weighted error to make the cost sensitive model. To improve the performance of the model, the origin-destination specific data is used along with all weather conditions, wind direction, speed etc. Based on null hypothesis, features are selected from the 50 features, only 20 features are considered for learning. Supervised learning models such as Decision tree, Random forest, AdaBoost and KNN algorithms are used. Experiment results showed that ensemble model AdaBoost has achieved the highest accuracy.

The work in [4] studied the ticket price and demand prediction from customer side as well as airline side. The paper discussed customer side prediction as two phases optimal ticket purchase time prediction and predict the exact price of ticket. Feature selection techniques for price prediction is discussed along the model used for prediction including Neural Network, Nash equilibrium, and trigonometric functions. The paper concluded that this area of research is less studied and has more scope to work.

**INFERENCE FROM EXISTING STUDY**

It is observed that all the existing work researched more for flight time delay prediction in airlines. Most of work used machine learning algorithm for prediction. Airline prediction based on customer review and rating is done attempted in earlier work. One more problem is that, all existing work studied under machine learning and not used deep learning algorithm. Thus we interested to implement a deep learning model.

**CHAPTER 3**

**SYSTEM ANALYSIS**

**Motivation of the project**

The primary motivation of the project is to provide a good travel experience to the customers, the customer can choose the airline with good rating and reviews, thus automating this process with machine learning or deep learning model will aid the customers to choose for the right airline services. Using this application, user can avoid any disappointment in the services and plan for a cost effective travel. The other motivation of the project is to enhance the customer satisfaction based on review or rating analysis, this will increase the business for airline service providers to excel their services in the lagging areas.

**Interests/ Computer science principles**

This project is selected on interest that the airline prediction is less attempted in the research. The real time data is web scrapped and used for this application is more advantageous to predict the airline based on the customer ratings or reviews.

The computer science principles used on this project includes classification problem attempted using machine learning and deep learning approaches. Create a web application using Html and flask application for users to help in choosing the appropriate airlines.

**Purpose of Project**

The purpose of this project, airline prediction system based on the customer reviews or rating provided for each airline. The purpose of the project is to design an application, which helps customers to choose source and destination and get the airlines based on the predictions. This is achieved by implementing machine learning AdaBoost and deep learning algorithm CNN. The proposed work considered dataset for rating with some of the important attributes such as seat comfort level, cabin staff service, food and beverages, etc. Similarly, for reviews, the text reviews written by the customers are processed with Natural language processing. Thus the purpose of projects as given below

* Based on Reviews (Using NLP) - Predict airline with AdaBoost and CNN models
* Based on ratings - Predict airline with AdaBoost and CNN models

**Existing Work**

* In the existing works, online reviews are analyzed for fake or real using machine learning algorithms in online ecommerce portals.
* Existing work analyzed reviews and based the reviews recommendations are available for online shopping websites.
* Existing works are analyzed on flight delay prediction using machine learning are available.
* No research on airline prediction is available.

**Proposed work**

* The proposed system is real time airline prediction based on the customer reviews or rating provided for airline.
* The dataset is considered for rating with some of the important attributes such as seat comfort level, cabin staff service, food and beverages, etc. Similarly, for reviews, the text review written by the customers are processed with Natural language processing.
* Proposed system used machine learning algorithm AdaBoost and deep learning algorithm CNN for implementation on rating and review based dataset.

**CHAPTER 4**

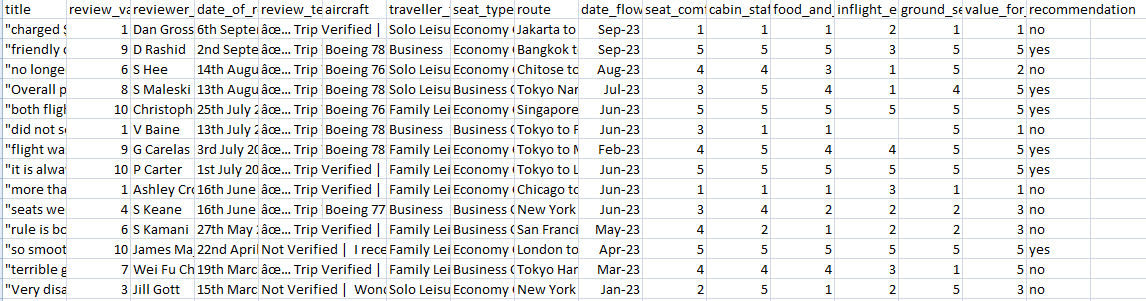
**METHODOLOGY**

The methodology of this project follows the following modules of implementation

1. Real time data collection from airlines website
2. Data cleaning and data pre-processing
3. Web UI with flask application
4. Review based implementation of Ensemble Machine learning
5. Rating based implementation of Ensemble Machine learning
6. Review based implementation of CNN
7. Rating based implementation of CNN

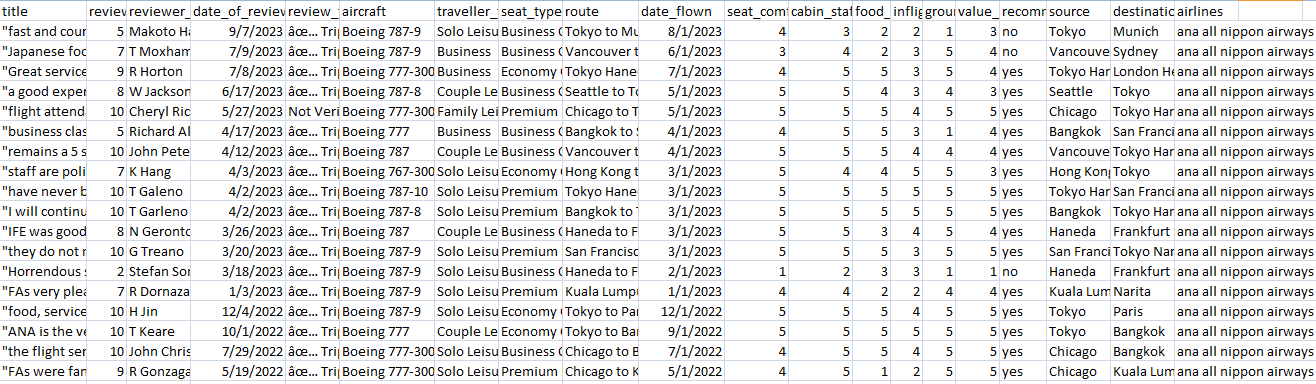
**1. Real time data collection from Airlines website**

The dataset is web scrapped from airlinequality.com. The data downloaded contains with different airlines name (Example: japan-airlines, lufthansa, qatar-airways, etc). The different airlines data are web scrapped and stored as csv files. The data from each file is finally concatenated to get a complete dataset file, which is used for learning. Python library BeautifulSoup is used for parsing the data from html web pages. The below screen shows the data extracted for Japan airlines.



**Figure: Extracted data for 'japanairlines' with review and ratings**

Extracted data contained attributes are title, review value, reviewer name, date of review, review text, aircraft, traveler type, seat type, route, date flown, seat comfort rating, cabin staff, food and beverage rating, inflight entertainment rating, ground service rating, value for money rating, recommendation, source and destination and airlines. The below screen shows the concatenated data for all available airlines.



**Figure: Extracted and concatenated dataset with review and ratings for airlines**

The dataset as shown in the above screen is concatenated and used for data cleaning and pre-process for the next module.

**2. Data cleaning and data pre-processing**

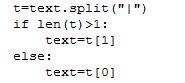
The extracted data from airlinequality.com needs some pre-processing, below are the steps involved in data cleaning and processing. There are some columns with null values are pre-processed by filling the null values to '0'

***Cleaning Review column:***

The review text contained unwanted text is removed, example review as given below

âœ… Trip Verified | I was charged $200 for my surfboard. Rude customer service and no option to change to an exit row. Overall a bad experience.

Here the first few characters are removed by splitting using |. The code snippet is given below



The above code omits the data till it find '|' and considers the remaining part of the text, if the splitter not found, it simply takes the text.

Converting the text to lower case is performed and stop word removal is done. The next step of pre-processing is stemming is performed to convert the text to base form. Stemming helps in normalizing the text by reducing the text to base form. It helps in identifying the base form of text and further cleaned text can be used for vectorization.

***Vectorizing review column:***

The cleaned text data is applied Term Frequency- Inverse Frequency (TF-IDF) to vectorize. The method TfidfVectorizer is used Python NLTK package. TF-IDF is the computation of word frequency in the given document/text. TF is a measure of number occurrence of a word to the total number of words. IDF is the measure of total number of text/document to the total number of text/document containing the term. In this implementation, this method takes input as review text and converts it into numerical feature vectors, thus it can be used for learning. The code snippet is shown below.



***Categorizing the y-value:***

The target column / prediction attribute is 'recommendation', which is the column if users has recommended the airlines, then it is 'Yes', if they not recommend the airlines, it is 'No'. The values in the target column is categorized to numeric values. Thus 'no' is replaced as '0' and 'yes' is replaced as '1'. This classification, we considered as binary classification.

reviews['recommendation'] = reviews['recommendation'].replace(['no','yes'],[0,1])

***Label encoding:***

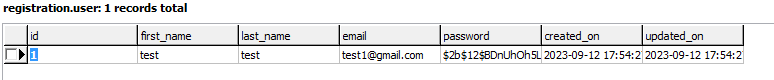
The attribute seat\_type has values such as Business class, Economy class, First class and premium. These text values are converted into numerical values by label encoder. Thus the seat\_type values will be converted into 0,1,2,3 for Business class, Economy class, First class and premium respectively.

reviews['seat\_type'] = labelencoder.fit\_transform(reviews['seat\_type'])

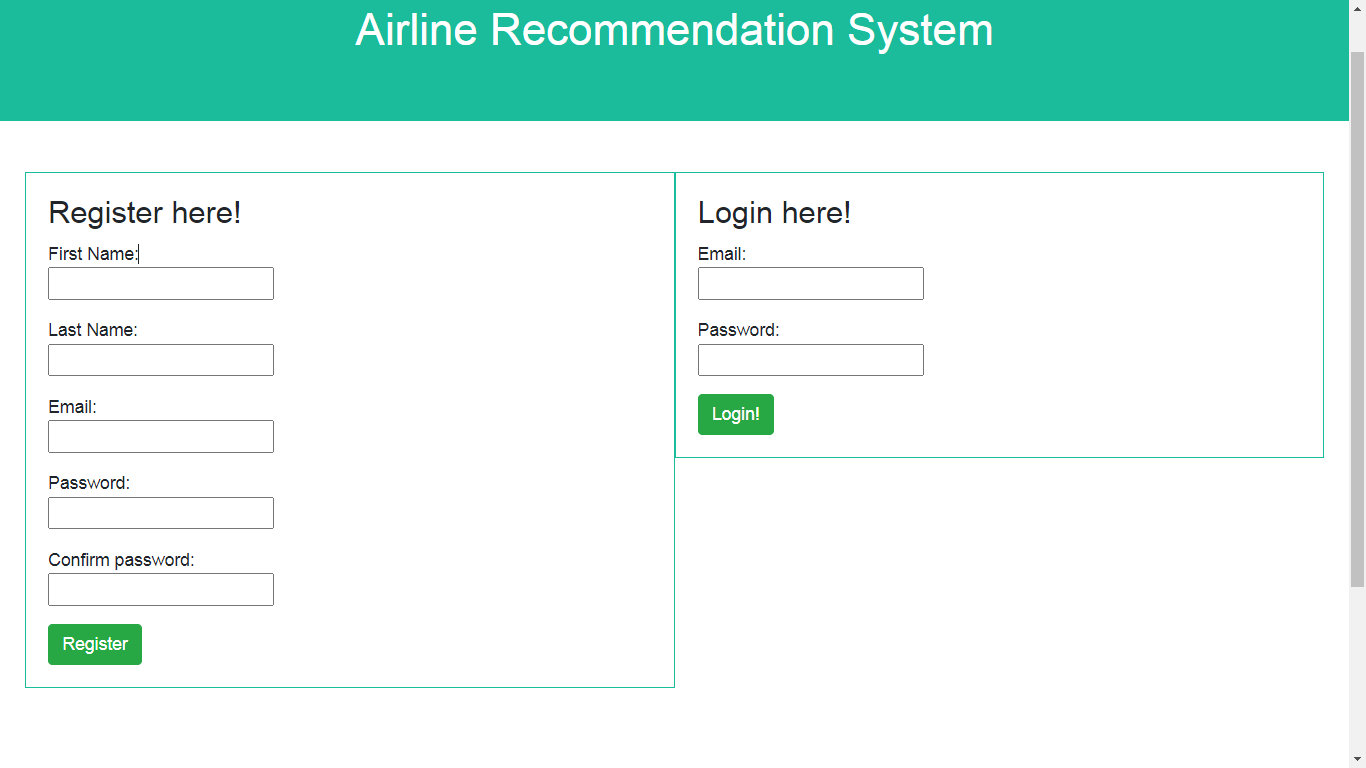
**3. Web UI on flask application**

We design user interface in HTML, flask application. The user interface has index page, where user can able to register and login. The database used for storing user data is MySQL database.

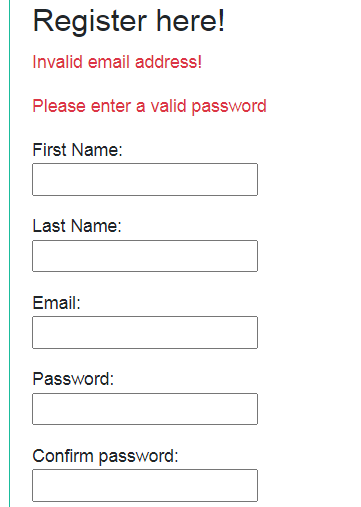
Database is stored with user information as shown below. The password is encrypted ans stored using the inbuilt function called bcrypt.generate\_password\_hash().



The following is user interface for user to login and access the airline prediction application.



The user registration and login are validated.



After login user can navigate to main page. The menu provided for creating dataset.

**Review based implementation of Ensemble Machine Learning (AdaBoost)**

The ensemble model used for this airline prediction is AdaBoost Model. AdaBoost is a supervised, boosting based ensemble model and is most preferred for binary classifications. As our Airline prediction is a binary classification, this ensemble model is preferred than other normal classification models. The cleaned dataset 'scraped\_reviews.csv' is taken with learning attribute (review\_text) and prediction attribute(recommendation) are taken. The dataset is cleaned text data with review text column is available for the implementation.

***Data Cleaning***

Review text is cleaned before converting it to vectorization data. The data cleaning involved splitting the data with delimiter (|) and omit the first part of data before delimiter and take the later part, which is the original review content taken for further cleaning. Punctuations in the text is removed and the text is converted to lower case. The stop words in English are removed from the review text as they add no meaning to the text processing. The stemmer is applied on the review text to get the base form of words, the stemmer applied here is snowball stemmer. The cleaned review text is applied vectorization technique.

***TF-IDF vectorization***

The cleaned review text data is applied term Frequency (TF) and Inverse Document Frequency (IDF) to convert the review text as vectorized data. This technique TF-IDF converts the text based on its count present in the review, the formula is given as below.

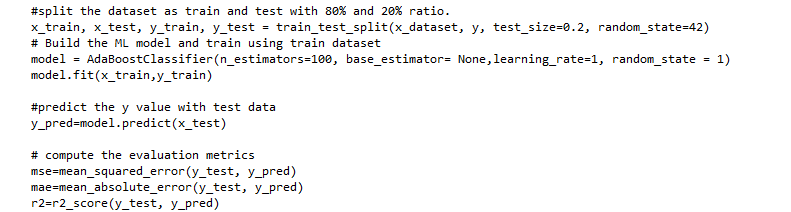
TF= number of times the term occurs in review/total number of terms in the dataset --(1)

IDF = log(total number of reviews in dataset/number of reviews containing the term)---(2)

Based on the above equations (1) and (2) the TF IDF constructs the vectorized data for the review content.

***AdaBoost classification***

The vectorized reviews are taken as X\_train and X\_text data, Y-values are recommendation provided in the dataset as 'Yes' and 'No' is converted to categorical numeric values '1' and '0'. AdaBoost algorithm is defined with number of estimators 100, base estimator None, learning rate is 1. The model is fit with X\_train and Y\_train values. Then predicted using X\_test value. The dataset is split as 80% training data and 20% test dataset.



**Figure: Code snippet for AdaBoost model- Review based Implementation**

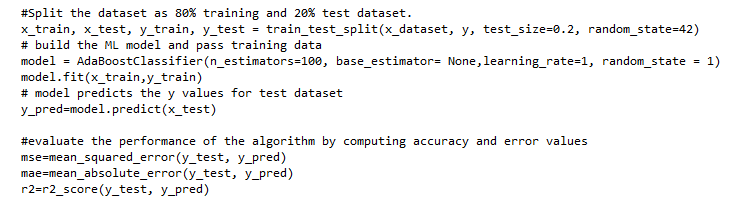
The following figure shows the read the dataset from csv file and loaded to model.



**Figure: Loading data after checking the null data**

**Rating based implementation of Ensemble Machine Learning (AdaBoost)**

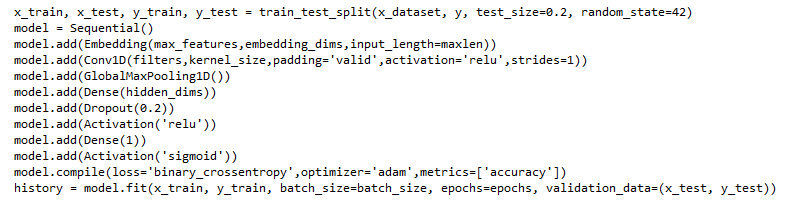
The cleaned dataset form the previous model is taken as input for this module. The rating based analysis included the attributes such as review\_value, seat\_comfort\_rating, cabin\_staff\_service\_rating, food\_and\_beverages\_rating, inflight\_entertainment\_rating, ground\_service\_rating, value\_for\_money\_rating, recommendation. The dataset is checked for null values and replaced null with 0. AdaBoost algorithm is defined with number of estimators 100, base estimator None, learning rate is 1. The model is fit with X\_train and Y\_train values. Then predicted using X\_test value. The dataset is split as 80% training data and 20% test dataset.



**Figure: Code snippet for AdaBoost model -Rating based Implementation**

**Review based Implementation of CNN**

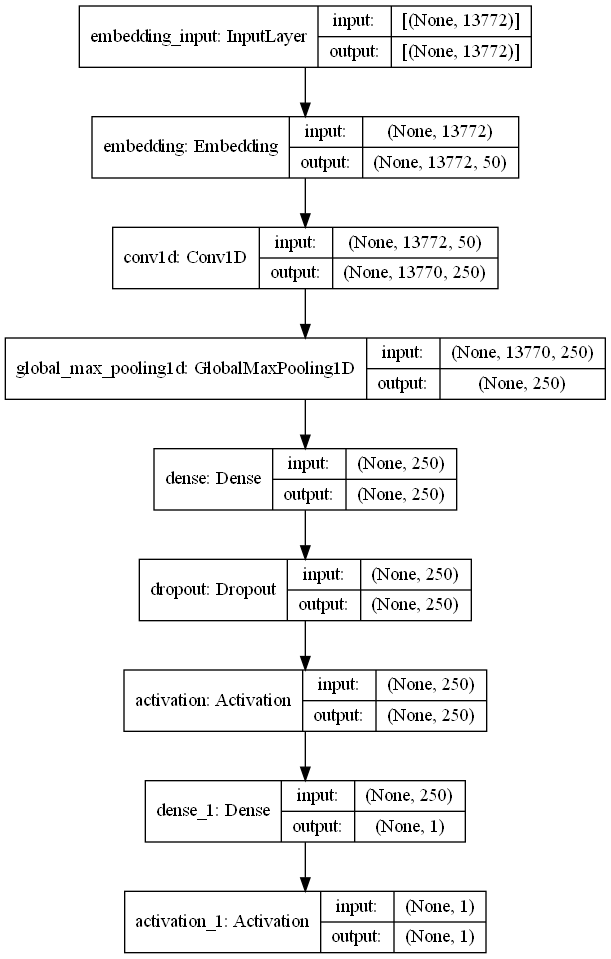
Convolutional Neural Networks are preferred deep learning model as it can extract the important features in dataset. In this work, we used CNN model to predict the airline based on reviews. The necessary pre-process required for reviews as it has text content. The data cleaning followed for AdaBoost review model is followed here. The CNN model is built with input layer by giving input dimension as max\_features, embedding dimension, input length. The model is activated by sequential model. The input layer is added as an embedding layer, which converts integer encoded text to dense vector having size 50 (embedding dimension). Followed by input layer, a convolutional 1D layer is added with kernel size 3, 250 filters, activation unit 'Rectified Linear Unit' relu is used. Followed by this, max pooling layer is added, which is used for dimensionality reduction. A dense hidden layer is added with 250 filter size and dropout of 20% is added. Another Dense layer is added, which outputs single unit with sigmoid activation, which gives the output of binary classification results. The model is compiled with 3 epochs, batch size used is 32, the model is compiled with categorical cross entropy loss and accuracy metrics are evaluated.



**Figure: Code snippet for CNN model**

**Rating based implementation of CNN**

CNN model architecture for rating based implementation is same as the review based implementation. The architecture of CNN used is represented below.



**Figure: Architecture of CNN model used for Review/Rating based implementations**

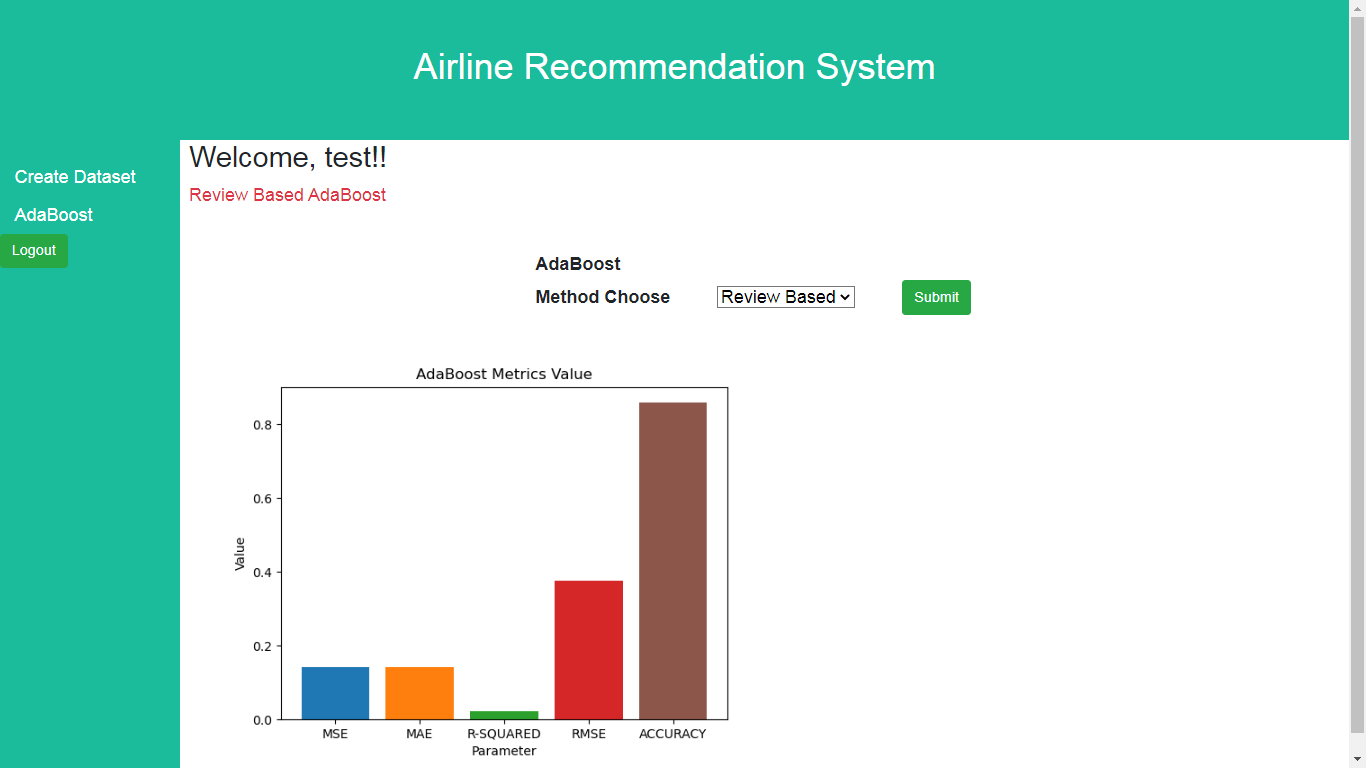
**CHAPTER 5**

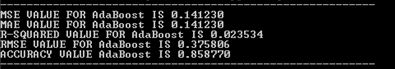
**EXPERIMENTAL RESULTS AND ANALYSIS**

Airline recommendation prediction using machine learning and deep learning model is proposed. The dataset is web scrapped from airlinequality.com. The data downloaded contains binary classes of recommendation as attributes to predict as 'yes' or 'no'. The prediction attribute is recommendation. The dataset is split into training and test data as 80% and 20% respectively. 80% of the data is used for training the algorithm and 20% is used for evaluation. The metrics used for evaluation includes Mean absolute error, Mean square error and accuracy for AdaBoost and Loss & accuracy metrics for CNN algorithm.

**Review Based AdaBoost model**

Review based learning model of AdaBoost has achieved the accuracy of 85.87%, the accuracy and error metrics plot is provided below.

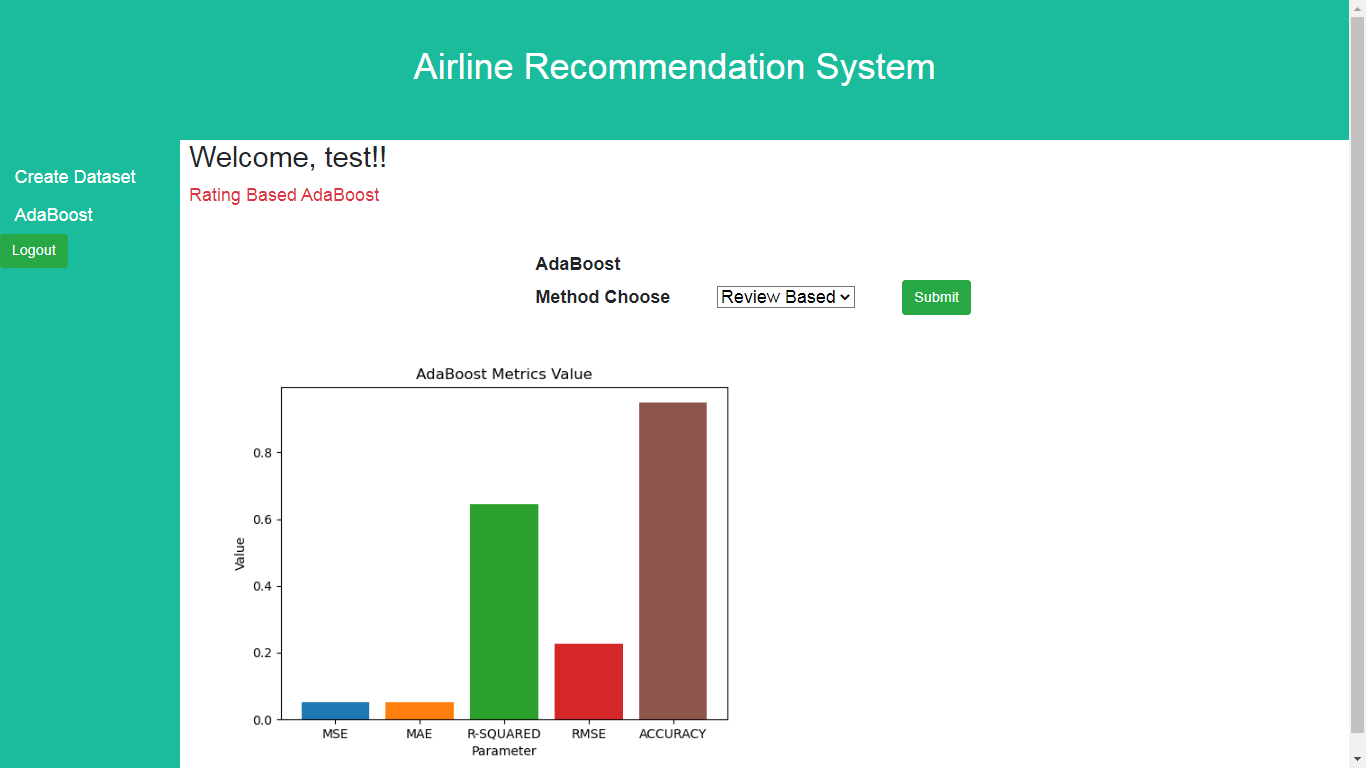
**Figure: Evaluation Results of AdaBoost Algorithm for Review based Airline recommendation**

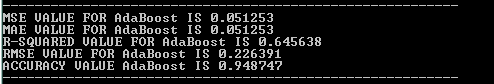


The above results shows that the AdaBoost algorithm performed well for review based airline prediction. Mean square error (MSE) value is 0.1412, Mean Absolute error (MAE) value is 0.1412, R square (R2) value is 0.0235, Root Mean Square error (RMSE) value is 0.3758. Accuracy of the AdaBoost algorithm for airline prediction is 85.88%.

**Rating Based AdaBoost model**

The below screen shows the results of rating based airline prediction. Accuracy of AdaBoost model for rating based airline recommendation is 94.87%.





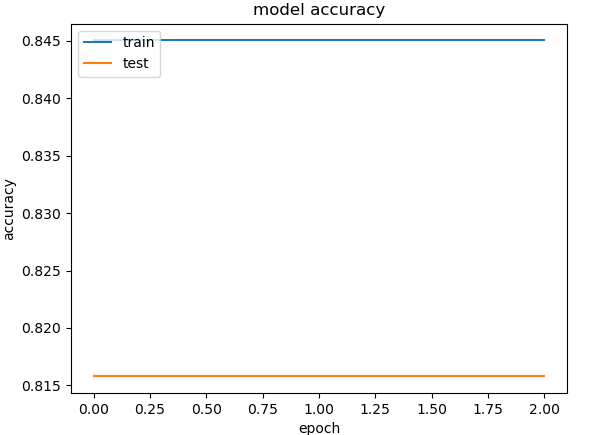
**Figure: Results of AdaBoost Algorithm for Rating based Airline prediction**

The results shows that the algorithm is performed better with rating based prediction. The above results shows that Mean square error (MSE) value is 0.0512, Mean Absolute error (MAE) value is 0.0512, R square (R2) value is 0.6456, Root Mean Square error (RMSE) value is 0.2263. Accuracy of the AdaBoost algorithm for airline prediction is 94.87%.

The results of Adaboost algorithm for the implementation of review based airline prediction and rating based airline prediction shows that the rating based prediction has achieved the highest accuracy of 94.87%.

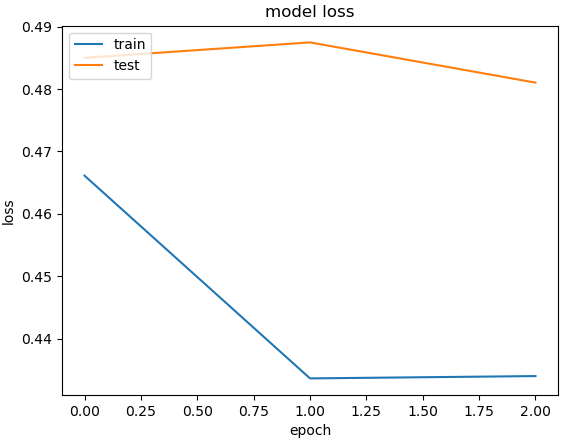
**Review Based CNN model**

Review based learning model of CNN has achieved the accuracy of 95%, the accuracy plot is shown below. The plot shows that the CNN training accuracy is around 84.5%. The validation accuracy of CNN model for review based model is 81.7%.



**Figure: Accuracy of CNN Algorithm for Review based Airline recommendation**

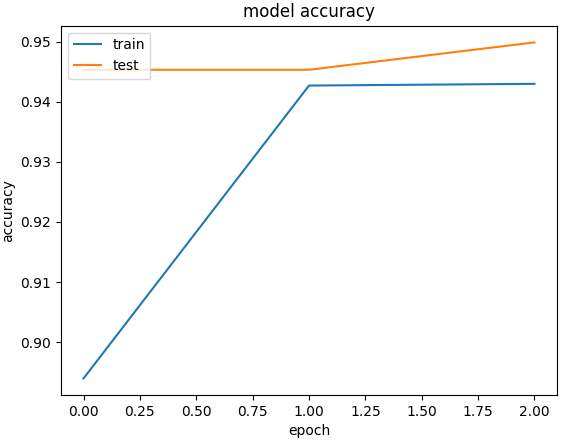
The below plot shows the loss metric for review based model. The training loss is around 0.44 and validation loss is 0.49.

****

**Figure: Loss metrics of CNN Algorithm for Review based Airline recommendation**

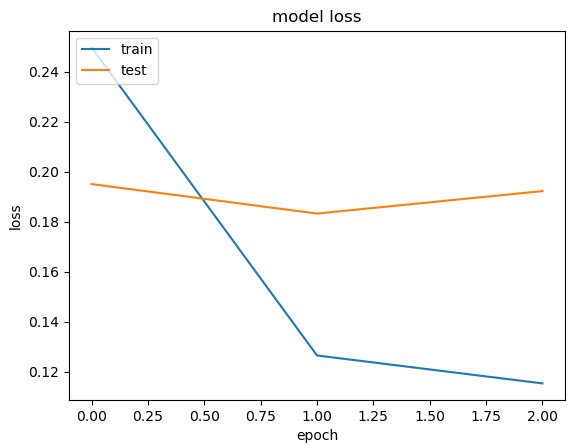
**Rating Based CNN model**

Rating based learning model of CNN has achieved the accuracy of 95%, the accuracy plot is shown below. Plot shows that the CNN model is more converged and the training accuracy achieved is around 94%. The validation accuracy of CNN model for review based model is 95%.



**Figure: Accuracy of CNN Algorithm for Rating based Airline recommendation**

The below plot shows the loss metrics of CNN for rating based model, the training loss is around 0.12 and validation loss is 0.19.



**Figure: Loss metric of CNN Algorithm for Rating based Airline recommendation**

The below table shows the comparison of performance of implemented models.

|  |  |  |
| --- | --- | --- |
|  | **Algorithm** | **Accuracy** |
| **Review Based Prediction** | AdaBoost | 85.88 |
| CNN | 81.70 |
| **Rating Based Prediction** | AdaBoost | 94.87 |
| CNN | **95.00** |

The experimental results showed that CNN rating based model has achieved the highest accuracy of 95%.

**CHAPTER 6**

**CONCLUSION AND FUTURE ENHANCEMENTS**

**CONCLUSION**

Airline recommendation prediction is challenging due to availability of data and multiple attributes considerations. The proposed project is based on machine learning and deep learning models to predict the recommendation. The users provided recommendation for airline based on the review and rating dataset is considered. The real-time data gathered from airlinequality.com. Machine learning based AdaBoost model and Deep learning, CNN model were implemented for review and rating based predictions. The experimental results showed that Rating based CNN model has achieved the highest accuracy of 95%.

**FUTURE ENHANCEMENTS**

As the further enhancement, we are interested to perform feature selection based on popular techniques like Genetic algorithm. We are also interested to implement more deep learning models like Deep Neural Network (DNN) and other pre-trained models such as DenseNet, AlexNet models.

**References**

[1] G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou and D. Zhao, "Flight Delay Prediction Based on Aviation Big Data and Machine Learning," in IEEE Transactions on Vehicular Technology, vol. 69, no. 1, pp. 140-150, Jan. 2020, doi: 10.1109/TVT.2019.2954094.

[2] S. Choi, Y. J. Kim, S. Briceno and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms," 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 2016, pp. 1-6, doi: 10.1109/DASC.2016.7777956.

[3] S. Choi, Y. J. Kim, S. Briceno and D. Mavris, "Cost-sensitive prediction of airline delays using machine learning," 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC), St. Petersburg, FL, USA, 2017, pp. 1-8, doi: 10.1109/DASC.2017.8102035.

[4] J. A. Abdella, N. Zaki and K. Shuaib, "Automatic Detection of Airline Ticket Price and Demand: A review," 2018 International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, 2018, pp. 169-174, doi: 10.1109/INNOVATIONS.2018.8606022.

[5] P. K. Jain, G. Srivastava, J. C. -W. Lin and R. Pamula, "Unscrambling Customer Recommendations: A Novel LSTM Ensemble Approach in Airline Recommendation Prediction Using Online Reviews," in IEEE Transactions on Computational Social Systems, vol. 9, no. 6, pp. 1777-1784, Dec. 2022, doi: 10.1109/TCSS.2022.3200890.

[6] Zhang, Yin, Rong Jin, and Zhi-Hua Zhou. ”Understanding bag-of words model: a statistical framework.” International Journal of Machine Learning and Cybernetics 1.1-4 (2010): 43-52.

[7] Qaiser, Shahzad, and Ramsha Ali. ”Text mining: use of TF-IDF to examine the relevance of words to documents.” International Journal of Computer Applications 181.1 (2018): 25-29.

[8] Zhang, Shichao, et al. ”Efficient knn classification with different numbers of nearest neighbors.” IEEE transactions on neural networks and learning systems 29.5 (2017): 1774-1785.