**Optimizing Flight Booking Decisions through Machine Learning Price Predictions**

**Mile stone 1 Define problem / problem understanding**

**Activity 1:**

***Business Problem***

Refer Project Description

**Activity 2:**

***Business requirement***

The business requirements for a machine learning model to predict personal loan approval include the ability to accurately predict loan approval based on applicant information, Minimise the number of false positives (approved loans that default) and false negatives (rejected loans that would have been successful). Provide an explanation for the model's decision, to comply with regulations and improve transparency.

**Activity 3:**

***Survey***

As the data is increasing daily due to digitization in the banking sector, people want to apply for loans through the internet. Machine Learning (ML), as a typical method for information investigation, has gotten more consideration increasingly. Individuals of various businesses are utilising ML calculations to take care of the issues dependent on their industry information. Banks are facing a significant problem in the approval of the loan. Daily there are so many applications that are challenging to manage by the bank employees, and also the chances of some mistakes are high. Most banks earn profit from the loan, but it is risky to choose deserving customers from the number of applications. There are various algorithms that have been used with varying levels of success. Logistic regression, decision tree, random forest, and neural networks have all been used and have been able to accurately predict loan defaults. Commonly used features in these studies include credit score, income, and employment history, sometimes also other features like age, occupation, and education level.

**Activity 4:**

***Impact***

Social Impact: -Personal loans can stimulate economic growth by providing individuals with the funds they need to make major purchases, start businesses, or invest in their education.

Business Model/Impact: - Personal loan providers may charge fees for services such as loan origination, processing, and late payments. Advertising the brand awareness and marketing to reach out to potential borrowers to generate revenue

Milestone 2 Data Collection & Preparation

ML depends heavily on data. It is the most crucial aspect that makes algorithm training possible. So this section allows you to download the required dataset.

Activity 1:

Collect the dataset

There are many popular open sources for collecting the data. Eg: kaggle.com, UCI repository, etc.

In this project we have used . csv data. This data is downloaded from kaggle.com. Please refer to the link given below to download the dataset.

Link: https://www.kaggle.com/code/anshigupta01/flight-price-prediction/data

As the dataset is downloaded. Let us read and understand the data properly with the help of some visualisation techniques and some analysing techniques.

Note: There are a number of techniques for understanding the data. But here we have used some of it. Inan additional way, you can use multiple techniques.

Activity 1.1: Importing the libraries

Import the necessary libraries as shown in the image. (optional) Here we have used visualisation style as fivethirtyeight.

Activity 1.2: Read the dataset

Our dataset format might be in .csv, excel files,.txt,.json, etc. We can read the dataset with the help of pandas.In pandas we have a function called read\_csv() to read the dataset. As a parameter we have to give the directory of csv file.

Activity 2: Data Preparation

As we have understood how the data is let’s pre-process the collected data. The download data set is not suitable for training the machine learning model as it might have so much of randomness so we need to clean the dataset properly in order to fetch good results. This activity includes the following steps.

• Handling missing values

• Handling categorical data

• Handling outliers

• Scaling Techniques

• Splitting dataset into training and test set

Note: These are the general steps of pre-processing the data before using it for machine learning. Depending on the condition of your dataset, you may or may not have to go through all these steps.

We have 1 missing value in Route column, and 1 missing value in Total stops column. We will meaningfully replace the missing values going further.

We now start exploring the columns available in our dataset. The first thing we do is to create a list of categorical columns, and check the unique values present in these columns

• 1. Airline column has 12 unique values - 'IndiGo' , 'Air India', 'Jet Airways' , 'SpiceJet' , 'Multiple carriers' , 'GoAir', 'Vistara', 'Air Asia', 'Vistara Premium economy' , 'Jet Airways Business', 'Multiple carriers Premium economy','Trujet'.

2. Source column has 5 unique values – ‘Bangalore’, ‘Kolkata’, ‘Chennai’, ‘Delhi’ and ‘Mumbai’.

3. Destination column has 6 unique values - 'New Delhi', 'Banglore', 'Cochin', 'Kolkata', 'Delhi' , 'Hyderabad'.

4. Additional info column has 10 unique values - 'No info', 'In-flight meal not included', 'No check-in baggage included', '1 Short layover' , 'No Info', '1 Long layover', 'Change airports' , 'Business class', 'Red-eye flight' , '2 Long layover'.

We now split the Date column to extract the ‘Date’, ‘Month’ and ‘Year’ values, and store them in new columns in our dataframe.

• Further, we split the Route column to create multiple columns with cities that the flight travels through. We check the maximum number of stops that a flight has, to confirm what should be the maximum number of cities in the longest route

• Since the maximum number of stops is 4, there should be maximum 6 cities in any particular route. We split the data in route column, and store all the city names in separate columns

• In the similar manner, we split the Dep\_time column, and create separate columns for departure hours and

minutes

• Further, for the arrival date and arrival time separation, we split the ‘Arrival\_Time’ column, and create

‘Arrival\_date’ column. We also split the time and divide it into ‘Arrival\_time\_hours’ and

‘Arrival\_time\_minutes’, similar to what we did with the ‘Dep\_time’ column

•Next, we divide the ‘Duration’ column to ‘Travel\_hours’ and ‘ Travel\_mins’

•We also treat the ‘Total\_stops’ column, and replace non-stop flights with 0 value and extract the integer part of the ‘Total\_Stops’ column

• We proceed further to the ‘Additional\_info’ column, where we observe that there are 2 categories signifying ‘No info’, which are divided into 2 categories since ‘I’ in ‘No Info’ is capital. We replace ‘No Info’ by ‘No info’ to merge it into a single category

We now drop all the columns from which we have extracted the useful information (original columns). We also drop some columns like ‘city4’,'city5' and ‘city6’, since majority of the data in these columns was NaN(null). As a result, we now obtain 20 different columns, which we will be feeding to our ML model. But first, we treat the missing values and explore the contents in the columns and its impact on the flight price, to separate a list of final set of columns.

• After dropping some columns, here we can see the meaningful columns to predict the flight price without the NaN values.

Activity 2.1: Replacing Missing Values

We further replace ‘NaN’ values in ‘City3’ with ‘None’, since rows where ‘City3’ is missing did not have any stop, just the source and the destination.

We also replace missing values in ‘Arrival\_date’ column with values in ‘Date’ column, since the missing values are those values where the flight took off and landed on the same date.

We also replace missing values in ‘Travel\_mins’ as 0, since the missing values represent that the travel time was in terms on hours only, and no additional minutes.

• Using the above steps, we were successfully able to treat all the missing values from our data. We again check the info in our data and find out that the dataset still has data types for multiple columns as ‘object’, where it should be ‘int’

•Hence, we try to change the data type of the required columns

• During this step, we face issue converting the ‘Travel\_hours’ column, saying that the column has data as ‘5m’, which is not allowing its conversion to ‘int’.

• The data signifies that the flight time is ‘5m’, which is obviously wrong as the plane cannot fly from BOMBAY->GOA->PUNE->HYDERABAD in 5 mins! (The flight has ‘Total\_stops’ as 2)

• We then convert the ‘Travel\_hours’ column to ‘int’ data type, and the operation happens successfully.

We now have a treated dataset with 10682 rows and 17 columns (16 independent and 1 dependent variable).

We create separate lists of categorical columns and numerical columns for plotting and analyzing the data

–

Activity 2.2: Label Encoding

• Label encoding converts the data in machine readable form, but it assigns a unique number (starting from 0) to each class of data. it performs the conversion of categorical data into numeric format.

• In our dataset I have converted these variables 'Airline','Source','Destination','Total\_Stops','City1','City2','City3','Additional\_Info' into number format. So that it helps the model in better understanding of the dataset and enables the model to learn more complex structures.

Activity 2.3: Output Columns

• Initially in our dataset we have 19 features. So, in that some features are not more important to get output (Price).

• So i removed some unrelated features and I selected important features. So, it makes easy to understand. Now we have only 12 Output Columns.

Milestone 3: Exploratory Data Analysis

Activity 1: Descriptive statistical

Descriptive analysis is to study the basic features of data with the statistical process. Here pandas has a worthy function called describe. With this describe function we can understand the unique, top and frequent values of categorical features. And we can find mean, std, min, max and percentile values of continuous features.

Activity 2: Visual Analysis

• Plotting countplots for categorical data

Activity 2.1: We now plot distribution plots to check the distribution in numerical data (Distribution of 'Price' Column)

• The seaborn.displot() function is used to plot the displot. The displot represents the univariate distribution of data variable as an argument and returns the plot with the density distribution. Here, I used distribution(displot) on 'Price' column.

• It estimates the probability of distribution of continous variable across various data.

Activity 2.2: Checking the Correlation Using HeatMap

• Here, I 'm finding the correlation using HeatMap. It visualizes the data in 2-D colored maps making use of color variations. It describes the relationship variables in form of colors instead of numbers it will be plotted on both axes.

• So, by this heatmap we found that correlation between 'Arrival\_date' and 'Date'. Remaining all columns don't have the any Correlation.

Activity 2.3: Outlier Detection for 'Price' Column

• Sometimes it's best to keep outliers in your data. it captures the valuable information and they can effect on statistical results and detect any errors in your statistical process. Here, we are checking Outliers in the 'Price' column.

Scaling the Data

• We are taking two variables 'x' and 'y' to split the dataset as train and test.

• On x variable, drop is passed with dropping the target variable. And on y target variable('Price') is passed.

• Scaling the features makes the flow of gradient descent smooth and helps algorithms quickly reach the minima of the cost function.

• Without scaling features, the algorithm maybe biased toward the feature which has values higher in

magnitude. it brings every feature in the same range and the model uses every feature wisely.

• We have popular techniques used to scale all the features but I used StandardScaler in which we transform the feature such that the changed features will have mean=0 and standard deviation=1.

Splitting data into train and test

Now let’s split the Dataset into train and test sets.

For splitting training and testing data we are using train\_test\_split() function. from sklearn. As parameters, we are passing x, y, test\_size, random\_state.

Milestone 4: Model Building

Now our data is cleaned and it’s time to build the model. We can train our data on different algorithms. For this project we are applying four regression algorithms. The best model is saved based on its performance.

Activity 1: Using Ensemble Techniques

RandomForestRegressor, GradientBoostingRegressor, AdaBoostRegressor

A function named RandomForest, GradientBoosting, AdaBoost is created and train and test data are passed as the parameters. Inside the function, RandomForest, GradientBoosting, AdaBoost algorithm is initialized and training data is passed to the model with .fit() function. Test data is predicted with .predict() function and saved in new variable. For evaluating the model, r2\_score, mean\_absolute\_error, and mean\_squared\_error report is done.

Activity 2: Regression Model

KNeighborsRegressor, SVR, DecisionTreeRegressor

A function named KNN, SVR, DecisionTree is created and train and test data are passed as the parameters. Inside the function, KNN, SVR, DecisionTree algorithm is initialized and training data is passed to the model with .fit() function. Test data is predicted with .predict() function and saved in new variable. For evaluating the model, r2\_score, mean\_absolute\_error, and mean\_squared\_error is done.

Activity 3: Checking Cross Validation for RandomForestRegressor

We perform the cross validation of our model to check if the model has any overfitting issue, by checking the ability of the model to make predictions on new data, using k-folds. We test the cross validation for Random forest and Gradient Boosting Regressor.

Activity 4: Hypertuning the model

RandomSearch CV is a technique used to validate the model with different parameter combinations, by creating a random of parameters and trying all the combinations to compare which combination gave them best results. We apply random search on our model.

From sklearn, cross\_val\_score is used to evaluate the score of the model. On the parameters, we have given rf (model name), x, y, cv (as 3 folds). Our model is performing well.

Now let’s see the performance of all the models and save the best model

Activity 5: Accuracy

Checking Train and Test Accuracy by Random SearchCV using Random ForestRegression Model Checking Train and Test Accuracy by Random SearchCV using KNN Model2 By Observing two models train and test accuracy we are getting good accuracy in Random ForestRegression

Activity 6: Evaluating performance of the model and saving the model

From sklearn, cross\_val\_score is used to evaluate the score of the model. On the parameters, we have given rfr (model name), x, y, cv (as 3 folds). Our model is performing well. So, we are saving the model by pickle.dump().

Note:To understand cross validation, refer this link. https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85.

Milestone 6: Model Deployment

Activity 1: Save the best model

Saving the best model after comparing its performance using different evaluation metrics means selecting the model with the highest performance and saving its weights and configuration. This can be useful in avoiding the need to retrain the model every time it is needed and also to be able to use it in the future.

Activity 2: Integrate with Web Framework

In this section, we will be building a web application that is integrated to the model we built. A UI is provided for the uses where he has to enter the values for predictions. The enter values are given to the, saved model and prediction is showcased on the UI.

This section has the following tasks

● Building HTML Pages

● Building server side script

● Run the web application

Activity 2.1: Building Html Pages:

For this project create two HTML files namely

● home.html

● predict.html

● submit.html

and save them in the templates folder.

Activity 2.2: Build Python code:

Import the libraries

Load the saved model. Importing the flask module in the project is mandatory. An object of Flask class is our WSGI application. Flask constructor takes the name of the current module (\_\_name\_\_) as argument.

Render HTML page:

Here we will be using a declared constructor to route to the HTML page which we have created earlier.

In the above example, ‘/’ URL is bound with the home.html function. Hence, when the home page of theweb server is opened in the browser, the html page will be rendered. Whenever you enter the values from the html page the values can be retrieved using POST Method.

Retrieves the value from UI:

Here we are rout ing our app to predict() function. This function retrieves all the values from the HTML page using Post request. That is stored in an array. This array is passed to the model.predict() function. This function returns the prediction. And this prediction value will be rendered to the text that we have mentioned in the submit.html page earlier.

Main Function:

Activity 2.3: Run the web application

● Open anaconda prompt from the start menu

● Navigate to the folder where your python script is.

● Now type “python app.py” command

● Navigate to the localhost where you can view your web page.

● Click on the predict button from the top left corner, enter the inputs, click on the submit button, and see the result/prediction on the web.

Now,Go the web browser and write the localhost url (http://127.0.0.1:5000) to get the below result Now, when you click on Predict button you will get redirected to the prediction page. Input 1- Now, the user will give inputs to get the predicted result after clicking onto the submit button.

Now when you click on submit button from right top corner you will get redirected to submit.html

Milestone 7: Project Demonstration & Documentation

Below mentioned deliverables to be submitted along with other deliverables

Activity 1:- Record explanation Video for project end to end solution

Activity 2:- Project Documentation-Step by step project development procedure

Create document as per the template provided