“FLIGHT PRICE PREDICTION PROJECT”

ACKNOWLEDGMENT

With great pleasure, I am sincerely express our deep sense everlasting profound gratitude and heartfelt thanks to several individuals from whom I received impetus motivation and invaluable work during the internship project work. I am very grateful to Keshav Bansal internship 17th batch guide, for the help and I do express our deep gratitude to him for able to guidance, keep interest and constant encouragement throughout our internship projects work. I am thanking her for personal concern, affinity and Great Spirit with which he has guided the work.

INTRODUCTION

1.Business Problem Framing

To be able to predict used cars market value can help both buyers and sellers.

There are lots of individuals who are interested in the used car market at some points in their life because they wanted to sell their car or buy a used car. In this process, it's a big corner to pay too much or sell less then it's market value. Machine Learning is a field of technology developing with immense abilities and applications in automating tasks, where neither human intervention is needed nor explicit programming.

The power of ML is such great that we can see its applications trending almost everywhere in our day-to-day lives. ML has solved many problems that existed earlier and have made businesses in the world progress to a great extent.

2.Conceptual Background of the Domain Problem

The project contains car price dataset. and we are supposed to predict the selling price of the car based on multiple features. we have used here random forest technique for selling price prediction. We are about to deploy an ML model for car selling price prediction and analysis. This kind of system becomes handy for many people.

Imagine a situation where you have an old car and want to sell it. You may of course approach an agent for this and find the market price, but later may have to pay pocket money for his service in selling your car. But what if you can know your car selling price without the

intervention of an agent. Or if you are an agent, definitely this will make your work easier.

Yes, this system has already learned about previous selling prices over years of various cars.

3.Review of Literature

All possible information from all the available data tables more the information, more than for EDA and feature Engineering. its take more important to take the average during the aggregation of data from tans the table rather than taking the counts before the loan was applies no future information steps into the data to be used for modelling.

4.Motivation for the Problem Undertaken

Any kind of modifications can also be later inbuilt in this application. It is only possible to later make a facility to find out buyers. This a good idea for a great project you can try out. You can deploy this as an app like OLA or any e-commerce app. The applications of Machine Learning don't end here. Similarly, there are infinite possibilities that you can explore. But for the time being, let me help you with building the model for flight Price Prediction and its deployment process.

Analytical Problem Framing

Mathematical/ Analytical Modelling of the Problem

1.Included in 'Data-cleaning ipynb'.

2 Selecting relevant features.

3 Exploratory the data analysis and data cleaning.

4 Null value imputation

5 Handling outliers

6 Training a machine learning model

7 Hyperparameter tunning

8 Evaluate the model

9 predictions of the model

1.Data Sources and their formats

Creating index on all the table to be joined this will speed up and processing here we are using EDA feature Engineering, Data visualization and statics approach perform data cleaning, outlier handling, missing values build model etc.

2.Data Pre-processing Done

Partitioning and splitting that dataset account when credit loan default was applied as that dataset is skewed, stratification is used allocate the samples evenly based on sample classes so that training set and test set have similar ration of classes.

3.Data Inputs- Logic- Output Relationships

Dataset is highly imbalanced we can use our models are majority and minority logic input and outputs.

Majority classes-the dataset is too small. down sampling the majority class will not help, so we will up sample the minority class.

Minority – we can balance the dataset either by up sampling the minority class or down sampling the majority class.

4.State the set of assumptions (if any) related to the problem under consideration

I am not taken any presumptions of this problem

5.Hardware and Software Requirements and Tools Used

1 processor= Intel i5 core 7^{th} gen

2 Motherhood-85EA

3 RAM – 8GB

4 keyboard version – 51.24

5 AMD Ryzen to keep temperature under control

6 smart Trooper cabinet

Packages

```
1  # Install python2 libraries
2  sudo apt-get install python
```

This project requires **Python 3.6.5** and the following Python libraries installed:

- Python 3.6.5
- NumPy
- Pandas
- matplotlib
- scikit-learn

You will also need to have software installed to run and execute a Jupyter Notebook

If you do not have Python installed yet, it is highly recommended that you install the Anaconda distribution of Python, which already has the above packages and more included.

Model/s Development and Evaluation

1.Identification of possible problem-solving approaches (methods)

1 examining the data

5 check the basic details (null value, D type, Shape etc)

5 identify the target and independent features and perform EDA using Data visualization and Statistical approach accordingly

5 perform data cleaning, outliers handling, missing value imputation

5 feature engineering

6 perform hyperparameter tunning

7 evaluate the model again

8 make prediction

2.Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing.

1 Decision Tree Regressor

5 Logistic Regression

5 Neighbours Regressor

5 SVM Regressor

5 Navie byes Regressor

6 Random Forest Regressor

3.Run and Evaluate selected models

1 cross validation
2 Hyperparameter tunning

5 evaluate the model

5 saving model
5 loading model
6 prediction

3.Key Metrics for success in solving problem under consideration

Skewness, removing outliers, classification methods

4.Visualizations

1 Sns.countplot(df[‘name’]); - here there is no columns all are equal.

5 Sns.countplot(df[‘date’]); -here all columns are not equal because of data imbalance.

5 sns.heatmap(cor,annot=True, linewidth=1,linecolor=’green’)-light shaded are highly correlated.

5 here all boxplots have outside the viscous we can easily find the outliers.

5 sns. Displot(df[‘name’])-here columns are normally distributed.

7 sns.displot(df[date’])-here columns are not normally distributed.

5.Interpretation of the Results

Visualization-explore the data, check the shape of data frame, check the co-relation between the features and remove the features which are highly co-related. Check the features for outliers, try to explore if the outliers are result of wrong entries or if they are import to the attribute under lens, outliers can lead to introduction of bias while training the model and can even lead to mis-classification.

Pre-processing-as the dataset is skewed, stratification is used allocate the samples evenly based on sample classes so that training set and test set have similar ratio of classes.

Modelling-running Decision Tree Regressor, Logistic Regression, SVM Regressor , Navie byes Regressor, Random Forest Regressor from sklearn package to get the Regression report and comparing the best models .use the Regression methods for choosing best model. Here the best Regression model is random forest Regressor

CONCLUSION :

1.Key Findings and Conclusions of the Study

Most Regression problems in the real world are imbalanced. Also, almost always data sets have missing values. Here covered strategies to deal with both missing values and imbalanced data sets. we also explored different ways of building ensembles in sklearn.

2.Learning Outcomes of the Study in respect of Data Science

There is no definitive guide of which algorithms to use given any situation. what may work on some data sets may not necessarily work on others. therefore, always evaluate methods using cross validation to get a reliable estimate.

sometimes we may be willing to give up some improvement to the model if that would increase the complexity much more than the percentage change in the improvement to the evaluation metrics.

In some Regression problems, false negative is a lot more expensive that false positives. Therefore, we can reduce cut-off points to reduce the false negative.

Missing values sometimes add more information to the model than we might expect.one way of capturing it is to add binary features for each feature that has missing values.

3.Limitations of this work and Scope for Future Work

1 Limited Access to information

2 Time Limits

3 conflicts on biased views and personal issues

4 How to structure my project research limitation correctly

5 How to set my project research limitation.

6 Formulation of my objectives and aims

7 Implementation of my data collection methods

8 Scope of discussions

9 Finding my error on the codes

10 Concluding thoughts.