**AUTOMOBILE PRICE PREDICTION**

# Technical Design Document

Version 1.0

RESTRICTED DISTRIBUTION

The information is standard Company Confidential, but due to its sensitivity, it has restricted distribution and viewing within iNeuron.

Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| Date Issued | Version | Description | Author |
| 25th June 2020 | 1.0 | Initial Draft | PRIYANK DUBEY  ATUL KUMAR |

Contributors

The content of this document has been authored with the combined input of the following group of key individuals.

|  |  |
| --- | --- |
| Name | Section Worked Upon |
| PRIYANK DUBEY  ATUL KUMAR | Initial Draft |

Document Classification

|  |  |
| --- | --- |
| Classification | Company Confidential |
| Definition | Information is Group confidential and needs to be protected |
| Context | Where the loss of information confidentiality would result in significant harm to the interests of the organisation, financial loss, embarrassment or loss of information |

Deployement of the model link :

<http://ec2-3-129-6-117.us-east-2.compute.amazonaws.com:8080/>

**CONTENTS**

1. INTRODUCTION

* CHALLENGE
* PROBLEM STATEMENT
* BUSINESS PROBLEM
* SOLUTION
* OBJECTIVES

1. WORKFLOW

* CHART

1. IMPORTANT PYTHON LIBRARIES USED IN THE PROJECT

* FOR DATA WRANGLING
* FOR DATA VISUALIZATION
* EDA
* MODEL BUILDING
* API DEVELOPMENT

1. STATS BASED EDA
2. graph based eda
3. STEPS
4. DATA CLEANING
5. FEATURE ENGINEERING AND FEATURE SELECTION
6. MODEL BUILDING
7. HYPERPARAMTER TUNING
8. MODEL SELECTION
9. UI DEVELOPMENT
10. DOCKERIZATION
11. DEPLOYMENT TO CLOUD
12. MODEL RETRAINING
13. MY HURDLES
14. MY LEARNING

# 1) INTRODUCTION:

CHALLENGE :

* What will be the price of this car ?

PROBLEM STATEMENT :

* To predict the car price on the basis of the given attributes of the car and the given features.

BUSINESS PROBLEM:

* In the business world it is manually very tough to predict a car price on the basis of given information of the sold cars and attributes.

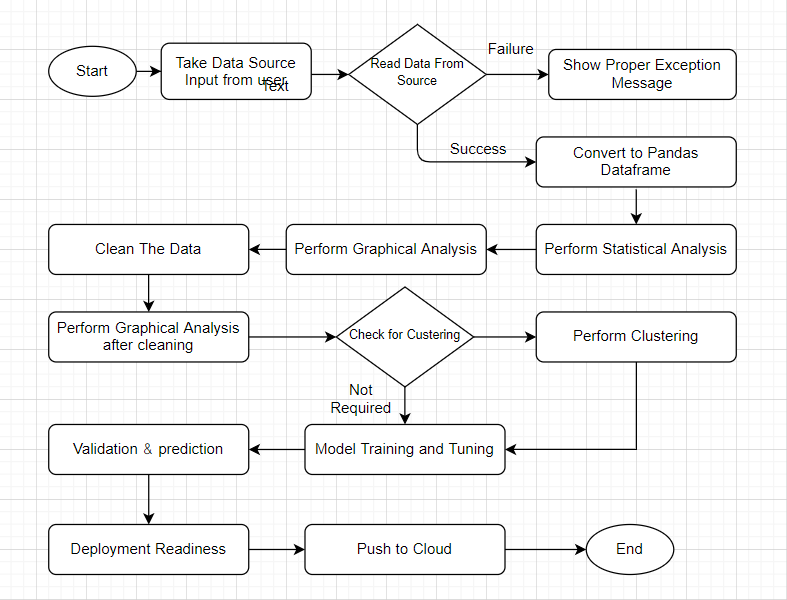
SOLUTION:

* Prediction of cars on the basis of the past data and featues becomes very easy and a minute work.

OBJECTIVES:

1. Enable reading/loading of data convert them into pandas dataframe.
2. Perform statistical analytics of the data and prepare a table for the analysis and show it on screen.
3. Perform graphical analysis for the data and Showcase the results (graphs) on the screen.
4. Perform data cleaning operation with all the steps required and showcase a report on screen.
5. After data cleaning showcase the graphical analysis once again for comparison.
6. Check whether clustering is required or not.
7. Choose the appropriate ML model for training.
8. Perform model Tuning.
9. Create a list of top 3 models and show multiple metrics for them.
10. Give option for prediction.
11. Give options for docker container creation.
12. Give option for automatic cloud deployment.

# 2) WORKFLOW:



3) IMPORTANT PYTHON LIBRARIES USED IN THE PROJECT:

**1) FOR DATA WRANGLING:**

* NUMPY
* PANDAS
* SIMPLEIMPUTER() CLASS OF SKLEARN

**2) FOR DATA VISUALIZATION:**

* SEABORN
* MATPLOTLIB
* PLOTLY

**3)EDA:**

* SWEETVIZ
* SCIPY
* STATSMODELS

**4)MODEL BUILDING:**

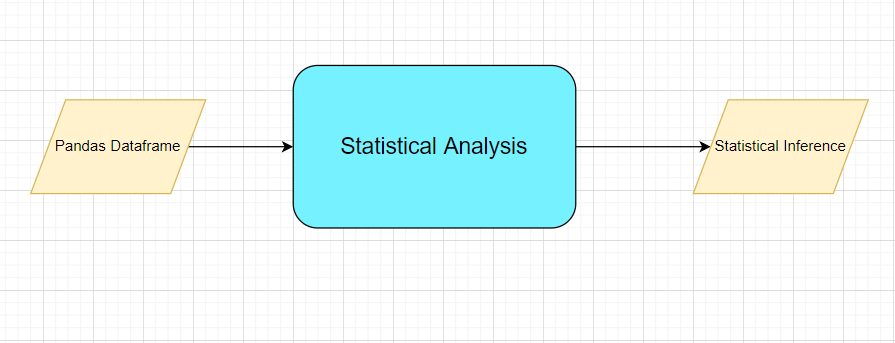
* XGBOOST
* SKLEARN

**5)API DEVELOPMENT:**

* FLASK
* REQUESTS

4)STATS BASED EDA:

* CORRELATION
* ANOVA TEST



5*)*GRAPHS USED ON EDA***:***

* DISTRIBUTION PLOTS FOR EACH NUMERICAL FEATURE
* PIE CHART FOR CATEGORICAL FEAUTRE
* BAR CHARTS
* HEATMAP FOR CORRELATIONS
* COUNT PLOTS
* BOX PLOTS

# 6) STEPS:

1) **DATA** **INFORMATION**:

The datset used in our model is from the UCI repository which is a automobiles (Car) dataset which has been recorded in the year 1985 , that means that the data set is very outdated as for now.Many features used in ther dataset are not even in use now-a-days .

Car price prediction is a regression problem which is our goal in the project so ,yes we have to predict the price of qa new car whose features are given:

1. **FEATURES** :

In our dataset we have a total of 205 records of different cars and columns count is 26 . There are many types of data ,some are nominal ,ordinal some are numerical .As we want to predict the price of the cars ,so price is our Target/response/output variable and all other features are independent features.

1. **Feature Information**:

Numerical: 15

Categorical :10

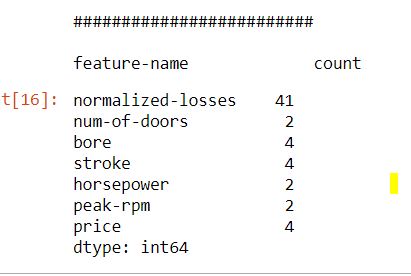
Target : 1

1. **Target variable**:

Price is our Target variable which is a continuous variable which is why our problem is a regression problem . In our data price value ranges from 5118 to 45000.

1. **Missing values**:

In our data as such there are not direct missing values but we have “?” in some columns which is then converted into Nan values and then the counts of missing values in our data is :



Except num-of-doors column all our columns are numerical so we have handled them differently.

1. For numerical columns we have replaced them into mean of them all records.
2. For our categorical features we have first found out the most occurred instant and then replaced all missing values with it .
3. And for our price column which is also our Target variable we have dropped all those rows which have missing values as it will not help us in our prediction goal.
4. **GATHERING INSIGHTS FROM THE DATA**:

After handling our missing values in the data we have moved to gather some more knowledge from the data .Like we have plotted count plot for our make feature in order to know which car company has the highest share in the cars dataset. Alse we have plotted pie charts for the Categorical features to know their share in the data .

# 7) DATA CLEANING:

Data cleaning is commonly defined as the process of detecting and correcting corrupt or inaccurate records from a dataset, table, or database.

Data quality is an important component in any data mining efforts. For this reason, many data scientists spend most of their time preparing and cleaning their data before it can be mined for insights.

1. **Outliers**:

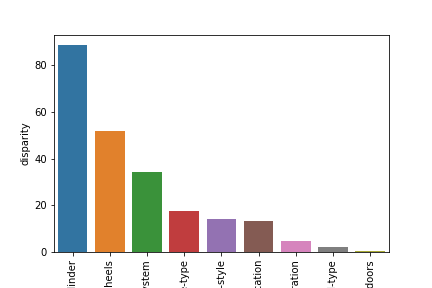
For this we have plotted first the box plots of all numerical columns and then observed that width and engine size column has the more number of outliers …..S o we have removed these outliers by the values of 75% quantiles of the respective feature.

Also in categorical columns we have some values which has only 1% of contributions so we have replaced them with the most occurred values .so their may not be imbalance in the data.

# 8) FEATURE ENGINEERING AND FEATURE SELECTION :

Feature engineering is a very mandatory steps in dat preparation for model . Features which have high signifance with the price variable are the good predictors of the price . In this step the features are standardized ,encoded and then finally good predictors of the target variable is chosen for further .

Features which are numerical are firstly scaled into a range with standardization and then the categorical features are encoded using proper encoders(OneHot,dummy variable, LabelEncoder). Then we have find out the correlation among target and indenependent features and then we have chosen the most significant features among the all using Pearsons correlation coefficient . For categorical features we have used Anova test to rank them as per their significance.



# 9) MODEL BUILDING :

Now we have our data prepared and we can now give green light to model building. Firstly we will split our data based on Training and Testing Data and then we will build the model on the Training and then validate it on Testing. Then will use xgboost regressor and Random forest regressor with Hyperparameter Tuning

To build our model.

1. RANDOMFOREST REGRESSOR

2) XGBOOST REGRESSOR

# **10)HYPER PARAMETER TUNING** :

In this steps we find the best parameters for the models to predict the best results .RandomizedCV and GridSearchCV is the best for hyperparameter tuning . In comparison Randomized CV

Is better because GridCV is less time effective . by providing the range for each parameter in the methods we can get the best ever parameters.

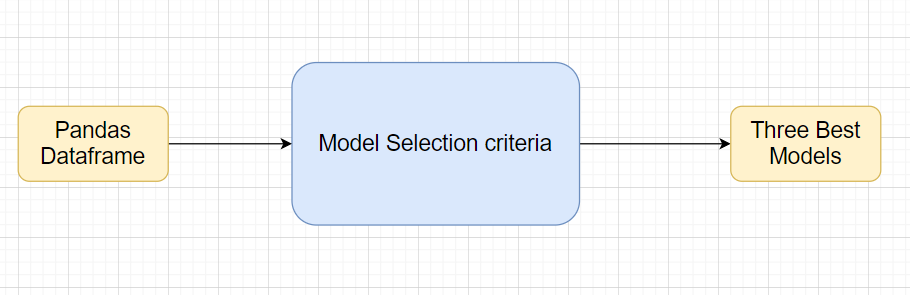
We have used the :

1) GridsearchCV with Xgboost

2) RandomizedCV with RandomForestRegressor

# 11) MODEL SELECTION :

We have actually used here two models , xgboost regressor and the RandomforestRegressor and then applied hyperparameter tuning in both of them and then calculated r2 values for both of them :



**Note:** The data should have been divided into train and validation set before this.

Methods for hyper tuning all kinds of models.

**Regression:**

Random Forest

XG Boost

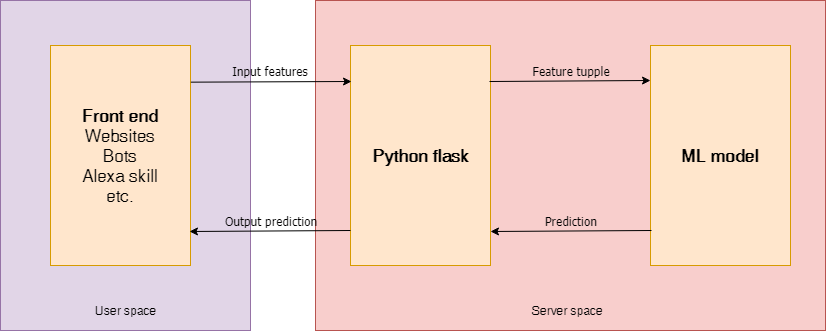
**MODEL SELECTION CRITERIA:**

R squared, adjusted R squared

-------🡪 From the first chart we can see that the results for xgboost and randomforesst are quite nice without tuninig itself

------🡪 After tuning also there results comes out to be nice because adjusted r2 is increased and coming equal to r2

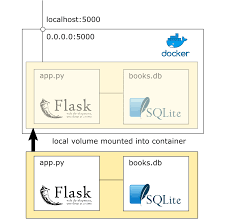
12) UI DEVELOPMENT:



Now comes the interface development part after model selection and in this we have used Flask to develop the interface where user will input the values of he feature and our model will provide the predictions on its basis and show the results.

13) DOCKERIZATION :

Now comes the docker part , **Dockerizing** an application is the process of converting an application to run within a Docker container, Making an application use environment variables when it relies on configuration files.

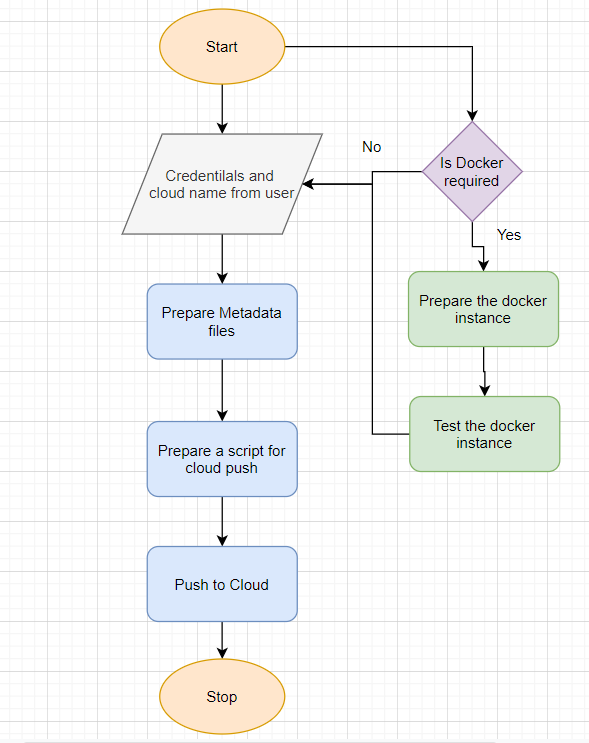


# 14) DEPLOYEMENT TO CLOUD :

Now we are on the final step of the project ,i.e cloud deployement . In this stage we will deploye the who;e docker imsge of our app to the cloud ,ie AWS,Azure,GCP.

We have deployed our app on aws ec2 instance using the Ubuntu server and the link of the deployed application is :

<http://ec2-3-129-6-117.us-east-2.compute.amazonaws.com:8080/>

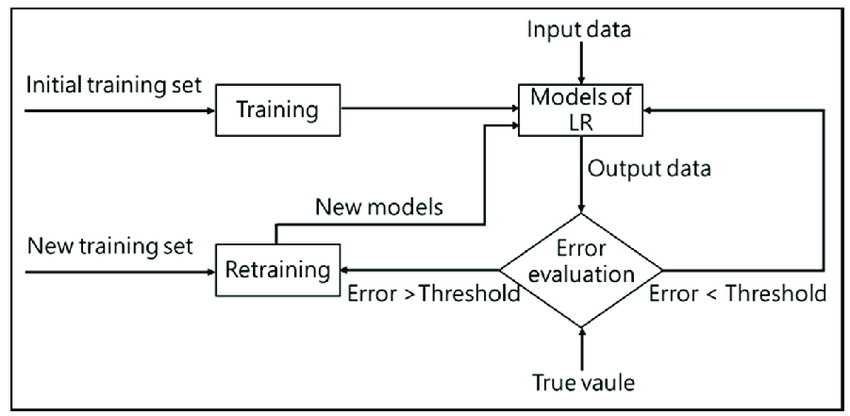


# 15) MODEL RETRAINING:

**Retraining** simply refers to re-running the process that generated the previously selected **model** on a new training set of data. The features, **model** algorithm, and hyperparameter search space should all remain the same. ... It only involves changing the training data set.

In That Model Ui Part We Have Alsoo Added An Optiion Of Model Retrainng In The Interfrace By Which A User Ca Retrain His Model On The Basis Of A New Data And Then On That BASIS OF MODEL FITTING THE USER CAN PREDICT FOR HIS TEST DATA.

Retraining w ill process the training set once again and rthen fit it into the best model and then load the best model .



# 16)MY HURDLES:

Following hurdles I have faced during the project:

1. The Data that has been given has a number of irrelevant features that have no significance in the model , maximum were categorical which were hard to be selected ,so it was a hurdle.
2. The dockerization part was very handy as it was performed by me for first time.
3. Cloud deployment on aws was also a hurdle because of its complexity.

# 17)LEARNING FROM THE PROJECT:

**FROM THE PROJECT I HAVE LEARNT THE FOLLOWING** :

1. Regresion ananlysis on the multivariate data using various algorithms ,using vrious methods and comparing the best out of them.
2. Building an api for the ml model for user friendly, so that a client can also undersatnd the model
3. Dockerizing the app to make i t work on any platform
4. And deploying to the cloud to make it public
5. Presenting our prooject in front of authorities with a presention and a document