CAR RESALE VALUE PREDICTION

## Category: Machine Learning

[IBM-EPBL](https://github.com/IBM-EPBL)/[**IBM-Project-27259-1660052392**](https://github.com/IBM-EPBL/IBM-Project-27259-1660052392)

**Team ID: PNT2022TMID06544**

**A PROJECT REPORT**

***Submitted by***

**YOGALINGAM M (19104091)**

**YEGAPPAN S (19104090)**

**VIGNESH M (19104085)**

**SANJAY PRASATH M (19104062)**

***in partial fulfillment for the award of the degree of***

**BACHELOR OF ENGINEERING**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**

**ERODE SENGUNTHAR ENGINEERING COLLEGE**

**(AUTONOMOUS)**

**TUDUPATHI-638057**

[INTRODUCTION 1](#_TOC_250014)

* 1. [Project Overview 1](#_TOC_250013)
  2. [Purpose 1](#_TOC_250012)

LITERATURE SURVEY 2

* 1. [Existing Problem 2](#_TOC_250011)
  2. Problem Definition 2
  3. References 3

[IDEATION AND PROPOSED SOLUTION 4](#_TOC_250010)

* 1. Empathy Map 4
  2. [Ideation and Brainstorming 4](#_TOC_250009)
  3. [Proposed Solution 5](#_TOC_250008)
  4. [Problem Solution Fit 6](#_TOC_250007)

REQUIREMENT ANALYSIS 7

* 1. Functional Requirements 7
  2. [Non-Functional Requirements 7](#_TOC_250006)

PROJECT DESIGN 8

* 1. Dataflow Diagram 8
  2. Technical Architecture 8
  3. [User Stories 9](#_TOC_250005)

[PROJECT PLANNING AND SCHEDULING 10](#_TOC_250004)

* 1. Sprint Planning & Estimation 10
  2. [Sprint Delivery Schedule 11](#_TOC_250003)
  3. Reports for JIRA 11

CODING AND SOLUTION 13

TESTING 14

* 1. [Test Cases 14](#_TOC_250002)
  2. [User Acceptance Testing 14](#_TOC_250001)

RESULTS 19

[9.1 Performance Metrics 19](#_TOC_250000)

ADVANTAGES AND DISADVANTAGES 16

CONCLUSION 16

FUTURE WORKS 17

APPENDIX 17

* 1. Source Code 19
  2. GitHub & Project Demo Link 19

**CHAPTER 1**

# INTRODUCTION

Given the variety of elements that influence a used car's market pricing, determining if the quoted price is accurate is a difficult undertaking. The goal of this research is to create machine learning models that can precisely forecast a used car's price based on its attributes so that buyers can make educated decisions. Implement and assess various learning techniques using a dataset of sale prices for various brands and models. The cost of the car will be determined based on a number of factors. Regression It is feasible to forecast the actual price of a car rather than just the price by using algorithms because they provide us a continuous value rather than a classified value as an output.

## Project Overview

* + - Able to understand the problem to classify if it is a regression or a classification kind of problem.
    - Able to know how to pre-process/clean the data using different data preprocessing techniques.
    - Applying different algorithms according to the dataset.
    - Able to know how to evaluate the model.
    - Able to build web applications using the Flask framework.

## Purpose

Car resale value prediction system is made with the purpose of predicting the correct valuation of used cars that helps users to sell thecar remotely with perfect valuation and without human intervention inthe process to eliminate biased valuation. Used car resale market in India was marked at 24.2 billion US dollars in 2019. Due to the huge requirement of used cars and lack of experts who can determine the correct valuation, there is an utmost need of bridging this gap betweensellers and buyers. This project focuses on building a system that can accurately predict a resale value of the car based on minimal features like kms driven, year of purchase etc. without manual or human interference and hence it remains unbiased.

1

## Existing problem

# CHAPTER 2 LITERATURE SURVEY

Using various machine learning algorithms we will predict the price. The algorithms involves Random Forest Regression. The best model which predicts the most accurate price is selected. After selection of the best model the predicted price is displayed to the user according to user’s inputs. User can give input through website to for used car price prediction to machine learning model.

## References

|  |  |  |
| --- | --- | --- |
| Book/journal | Author’s name | Inference |
| Predicting the Price of Used Cars using Machine Learning Techniques | Sameerchand Pudaruth | In this paper, we investigate the application of supervised machine learning techniques to predict the price of used cars in Mauritius.The predictions are based on historical data collected from daily newspapers. Different techniques like multiple linear regression analysis, k-nearest neighbour’s, naive bayesand decision trees have been used to make  the predictions. |
| Car Price Prediction Using Machine Learning | Enis gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric | In this paper, we applied different methods and techniques in order to achieve higher precision of the used car price prediction. This paper is organized in the following manner: Section II contains related work in the field of price prediction of used cars. In section III, the research methodology of our study is explain.  Section IV elaborates various machine learning. |
| Price Evaluation Model In Second Hand Car System Based On BP Neural Network Theory | Ning sun, Hongxi Bai, Yuxia Geng, Huizhu Sh | This paper presents a system that has been implemented to predict a fair price for any pre- owned car. The system works well to anticipate the price of used cars for the Mumbai region.  Ensemble techniques in  machine learning namely Random Forest Algorithm, extreme Gradient Boost. |
| Prediction of Prices for Used Car by using Regression Models | Nitis Monburinon, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, Pitchayakit Boonpou | In this paper, we look at how supervised machine learning techniques can be used to forecast car prices in India. Data from the online marketplace quikr was used to make the predictions. |

|  |  |  |
| --- | --- | --- |
| Prediction car prices using qualify qualitative data and knowledge-based system | Doan Van Thai, Luong Ngoc Son, Pham Vu Tien, Nguyen Nhat Anh, Nguyen Thi Ngoc Anh | In this paper, we describe a scalable end-to-end tree boosting system called XGBoost, which is used widely by data scientists to achieve state- of-the-art results on many machine learning challenges. We propose anovel sparsity-aware algorithm for sparse data and weighted  quantile sketch for approximate tree learning. |

* 1. Problem Statement Definition

The prices of new cars in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So customers buying a new car can be assured of the money they invest to be worthy. But due to the increased price of new cars and the incapability of customers to buy new cars due to the lack of funds, used cars sales are on a global increase. Predicting the prices of used cars is an interesting and much-needed problem to be addressed. Customers can be widely exploited by fixing unrealistic prices for the used cars and many falls into this trap. Therefore, rises an absolute necessity of a used car price prediction system to effectively determine the worthiness of the car using a variety of features. Due to the adverse pricing of cars and the nomadic nature of people in developed countries, the cars are mostly bought on a lease basis, where there is an agreement between the buyer and seller. These cars upon completion of the agreement are resold. So reselling has become an essential part of today’s world.

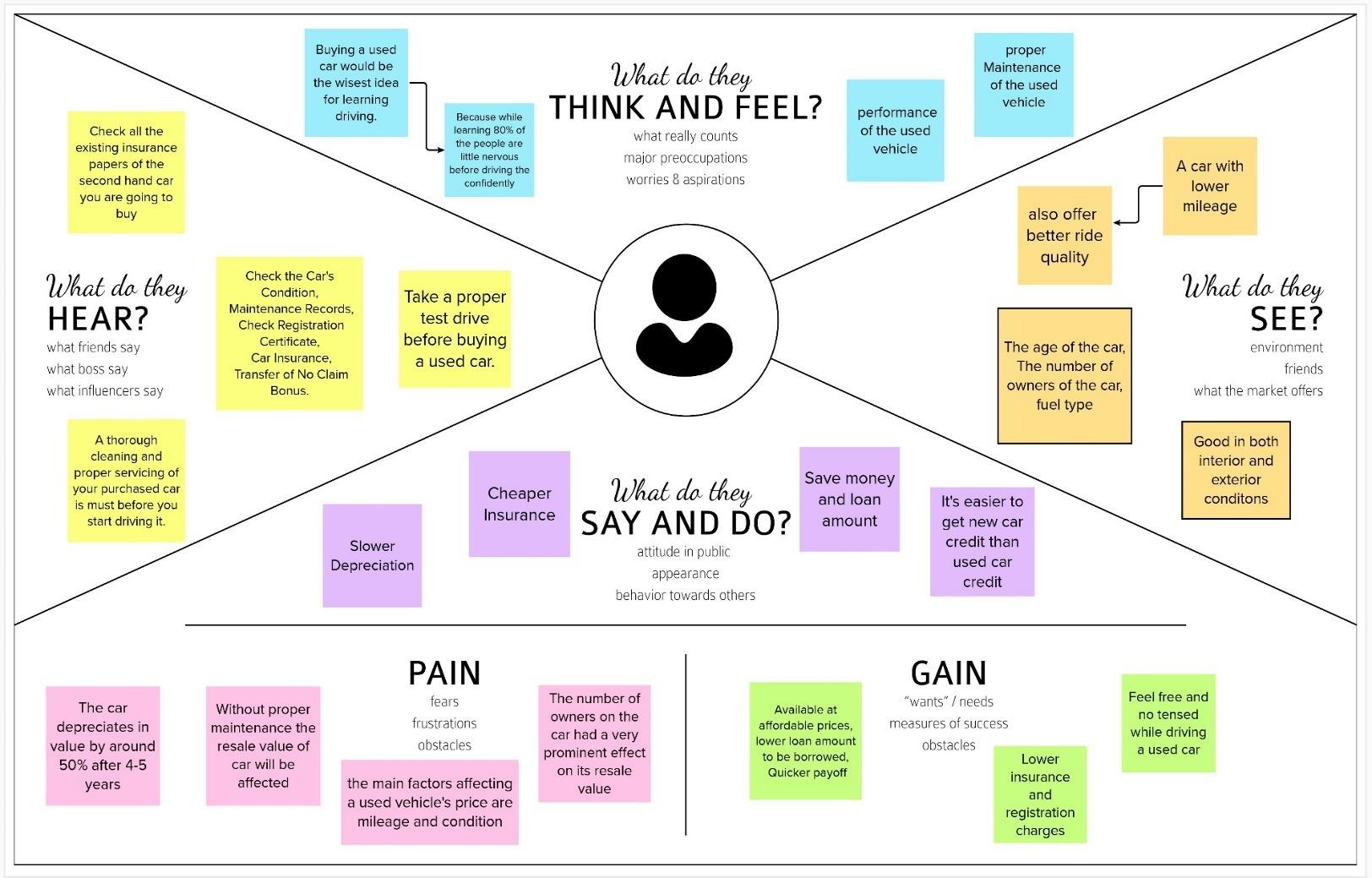
The main aim of this project is to predict the price of used cars using the various Machine Learning (ML) models. This can enable the customers to make decisions based on different inputs or factors namely Brand or Type of the car one prefers like Ford, Hyundai, Model of the car namely Ford Figo, Hyundai Creta, Year of manufacturing like 2020, 2021, Type of fuel namely Petrol, Diesel, Price range or Budget, Type of transmission which the customer prefers like Automatic or Manual, Mileage to name a few characteristic features required by the customer. This project Car Price Prediction deals with providing the solution to these problems. Different techniques like multiple linear regression analysis, k-nearest neighbours, naïve bayes and decision trees have been used to make the predictions. The predictions are then evaluated and compared in order to find those which provide the best performances.

3

# CHAPTER 3

# IDEATION AND PROPOSED SOLUTION

* 1. Empathy Map Canvas



## Ideation and Brainstorming

MEMBER 1:

* + - Car Prediction using image.
    - By using exterior and interior image of the car.
    - The value will be predicted based on the appearance of the car.If there any damage or n numbers scratches the car resale value will be quite affected.
    - By using neural network value of the car can be predicted
    - Neural network algorithm is developed by considering the human brain that takes a set of unit as input and transfers results to a predefined output

MEMBER2:

* + - The main objective of this project is to *predict* the *Prices* of *used cars*, compare the ***prices*** and also estimate the *life span* of a particular ***car.***
    - Insurance, Company claims,etc

4

* + - regression Algorithm is used to predict the value
    - Regression model based on k-nearest neighbor machine learning algorithm was used to predict the price of a car.

MEMBER 3:

* + - Car prediction using engine condition.
    - user should Upload engine Sound in the format of audio file.
    - By using Convolutional Neural Networks methodology price can be predicted.
    - CNNs for Machine Learning on sound data by spectrogram approach that was just converts each song (or song segment) into a spectrogram: a twodimensional matrix

MEMBER 4:

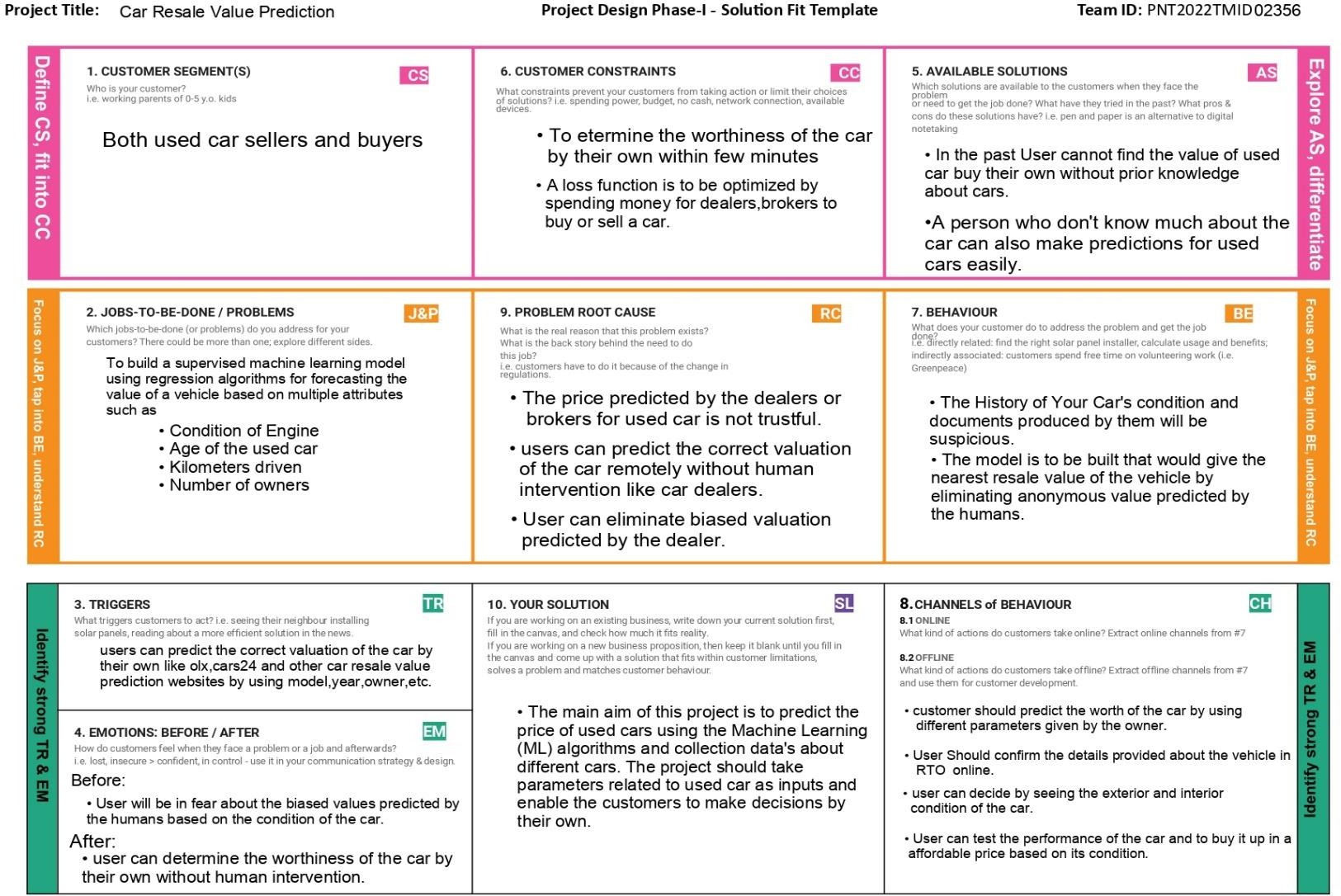
* + - Economic Conditions.
    - Kilometres Covered.
    - Its mileage (the number of kilometers it has run) and its horsepower
    - Car prediction using XGBoost algorithm accurate result will be monitored.
    - XGBoost as a regression model gave the best MSLE and RMSLE values.

## Proposed Solution

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Parameter** | **Description** |
| 1. | Problem Statement (Problem to be solved) | * The main aim of this project is to predictthe price of used cars using the various Machine Learning (ML) models. * The project should take parameters related to used car as inputs and enablethe customers to make decisions by theirown. |
| 2. | Idea / Solution description | * The model is to be built that would give the nearest resale value of the vehicle. By using these best accuracy value will be taken as a solution and it will be integrated to the web- based application where the user is notified with the status   of his product. |

|  |  |  |
| --- | --- | --- |
| 3. | Novelty / Uniqueness | * Used car price prediction is effectively used to determine the worthiness of thecar by their own within few minutes by using various features such as year, model, mileage(km), etc. |
| 4. | Social Impact / Customer Satisfaction | * If the user wants to buy or sell a own carit helps users to predict the correct valuation by their own. * A loss function is to be optimized and mainly a weak learner can make predictions for used cars easily. |
| 5. | Business Model (Revenue Model) | * It helps users to predict the correct valuation of the car remotely with perfect valuation and without humanintervention like car dealers in the process to eliminate biased valuation   predicted by the dealer. |
| 6. | Scalability of the Solution | * Using Stored data and machine learning approaches, this project proposed a scalable framework for predicting valuesfor different   type of used cars present allover India. |

## Problem Solution fit



# CHAPTER 4 REQUIREMENT ANALYSIS

* 1. Functional requirement

|  |  |  |
| --- | --- | --- |
| **FR No.** | **Functional Requirement (Epic)** | **Sub Requirement (Story / Sub-Task)** |
| FR-1 | User Registration | Registration through Gmail. |
| FR-2 | User Confirmation | Confirmation via Email. |
| FR-3 | User Login Window | Login using given credentials. |
| FR-4 | Dashboard | Fill the required data shown in the window. |
| FR-5 | Prediction Value | Predicting the value of the car and displayed in the window. |

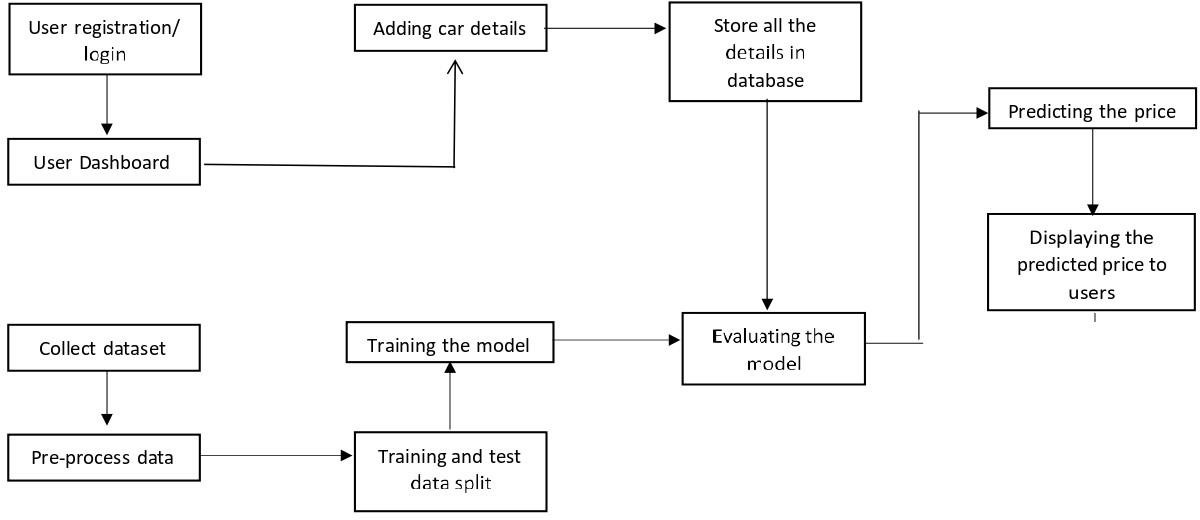
## Non-Functional requirements

|  |  |  |
| --- | --- | --- |
| **FR No.** | **Non-Functional Requirement** | **Description** |
| NFR-1 | **Usability** | Predicting the value of used cars. |
| NFR-2 | **Security** | Aware about fraudulent sites the data’s given by the user is not exposed in any way. |
| NFR-3 | **Reliability** | It helps user to predict the correct valuation of the car remotely with perfect valuation and without human intervention like car dealers. |
| NFR-4 | **Performance** | Users can determine the worthiness of the car by their own within a few minutes. |
| NFR-5 | **Availability** | It is available for everyone and can be accessed anywhere at anytime. |
| NFR-6 | **Scalability** | This project proposed a scalable framework for predicting values for different types of used cars  present all over India. |

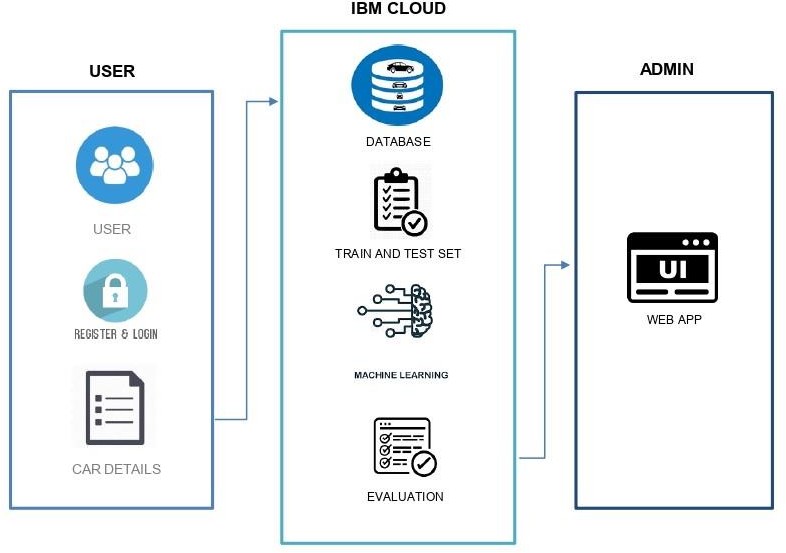
7

* 1. Data Flow Diagrams

# CHAPTER 5 PROJECT DESIGN



## Solution and Technical Architecture



8

## User Stories



9

# CHAPTER 6

# PROJECT PLANNING AND SCHEDULING

## Sprint Planning and Estimation

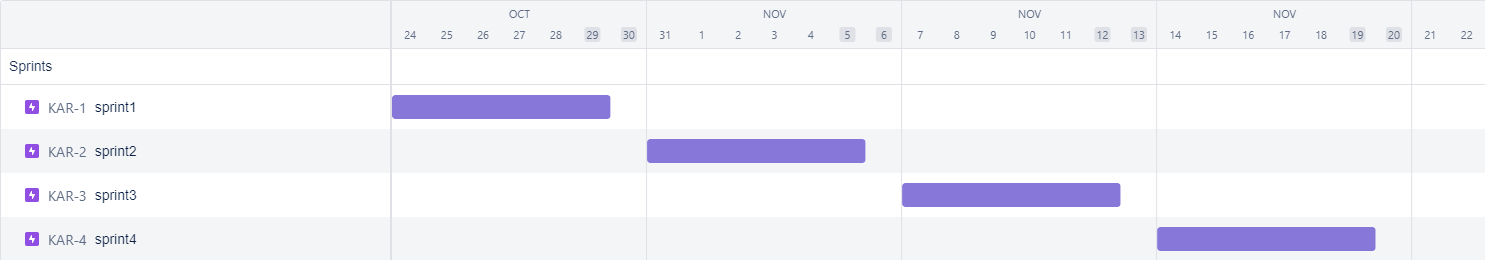
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **User** | **Sprint** | **Functional**  **Requirement (Epic)** | **User Story Number** | **User Story / Task** | **Story Points** | **Priority** | **Team Members** |
| Admin | Sprint 1 | Dataset collection | USN-1 | Collect the required data for the Car resaleprediction | 2 | High | YOGALINGAM M  YEGAPPAN S  VIGNESH M  SANJAY PRASATH M |
|  | Sprint 1 | Data pre- processing | USN-2 | Perform data cleaning to optimize the dataset | 4 | Medium | YOGALINGAM M  YEGAPPAN S |
|  | Sprint 1 | Training & Building Model | USN-3 | Build the model using regression algorithms toclassify the data | 6 | High | YOGALINGAM M  YEGAPPAN S  VIGNESH M  SANJAY PRASATH M |
|  | Sprint 2 | Deploy the model | USN-4 | Deployment of ML model using IBM Cloud | 5 | High | VIGNESH M  YOGALINGAM M |
|  | Sprint 4 | Integration | USN-5 | Integrate the web app developed using flask with  IBM model | 5 | High | YOGALINGAM M  YEGAPPAN S |
| Customer | Sprint 2 | Homepage | USN-6 | Details about the application and the car resale process | 2 | Low | YOGALINGAM M  SANJAY PRASATH M |
|  | Sprint 2 | Registration | USN-7 | As a user, I can register for the application by  entering confirming. | 5 | High | VIGNESH M  SANJAY PRASATH M  YEGAPPAN S |
|  | Sprint 3 | Confirmation | USN-8 | As a user, I will receive confirmation email once I  have registered for the application | 3 | Medium | YEGAPPAN S  SANJAY PRASATH M |
|  | Sprint 3 | Login | USN-9 | As a user, I can log into the application by entering email & password | 4 | High | YOGALINGAM M  VIGNESH M |
|  | Sprint 3 | Dashboard | USN-10 | As a user, I can add new cars and get access toinsert and update their details | 5 | High | YOGALINGAM M  VIGNESH M  YEGAPPAN S |
|  | Sprint 4 | Car Details | USN-11 | As a user, I should give the car details like car  model, engine and fuel type, etc… | 2 | Medium | VIGNESH M  SANJAY PRASATH M |
|  | Sprint 4 | Car Price | USN-12 | As a user, I can view the current rate of the usedcar price | 5 | High | YOGALINGAM M  YEGAPPAN S  VIGNESH M  SANJAY PRASATH M |

10

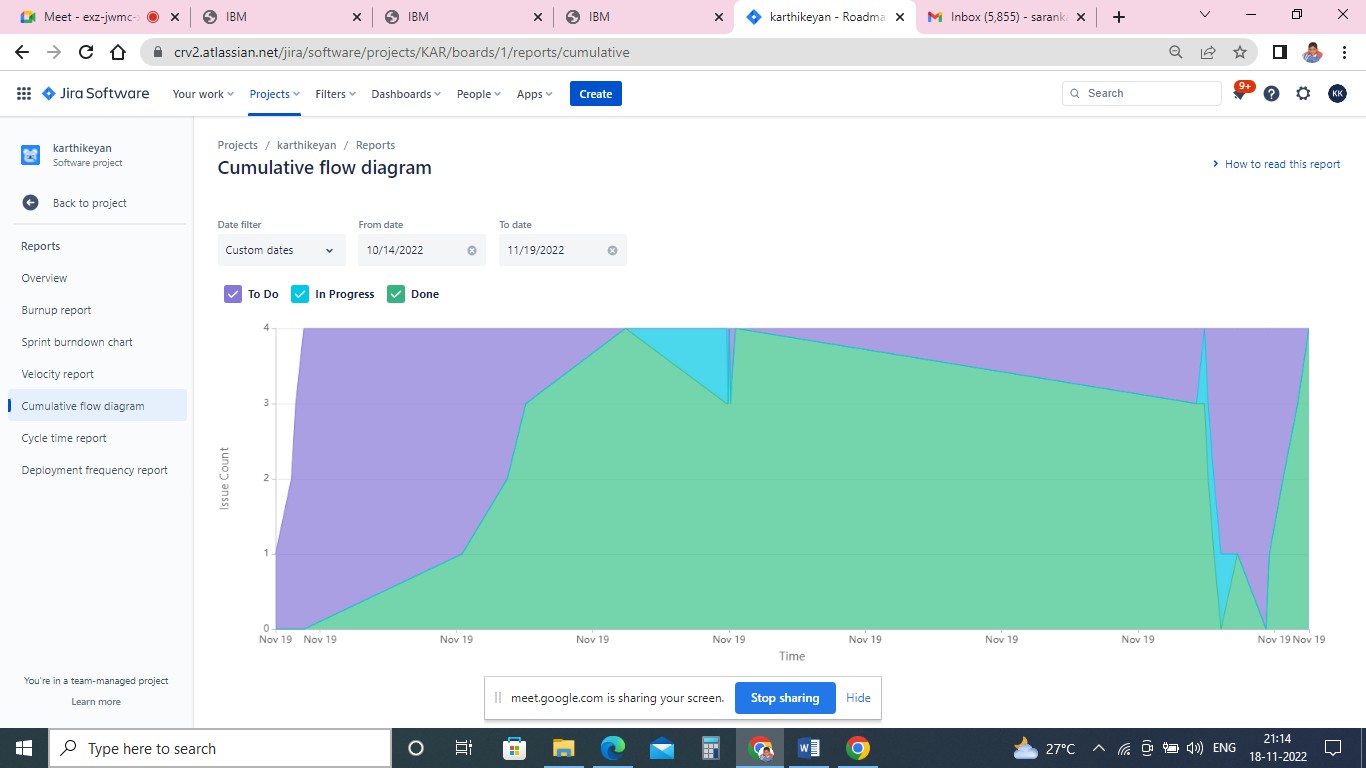
## Sprint Delivery Schedule

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sprint** | **Total Story Points** | **Duration** | **Sprint Start Date** | **Sprint End Date (Planned)** | **Story Points Completed (as on Planned End Date)** | **Sprint Release Date**  **(Actual)** |
| Sprint-1 | 12 | 6 Days | 24 Oct 2022 | 29 Oct 2022 | 12 | 29 Oct 2022 |
| Sprint-2 | 12 | 6 Days | 31 Oct 2022 | 05 Nov 2022 | 12 | 05 Nov 2022 |
| Sprint-3 | 12 | 6 Days | 07 Nov 2022 | 12 Nov 2022 | 12 | 12 Nov 2022 |
| Sprint-4 | 12 | 6 Days | 14 Nov 2022 | 19 Nov 2022 | 12 | 19 Nov 2022 |

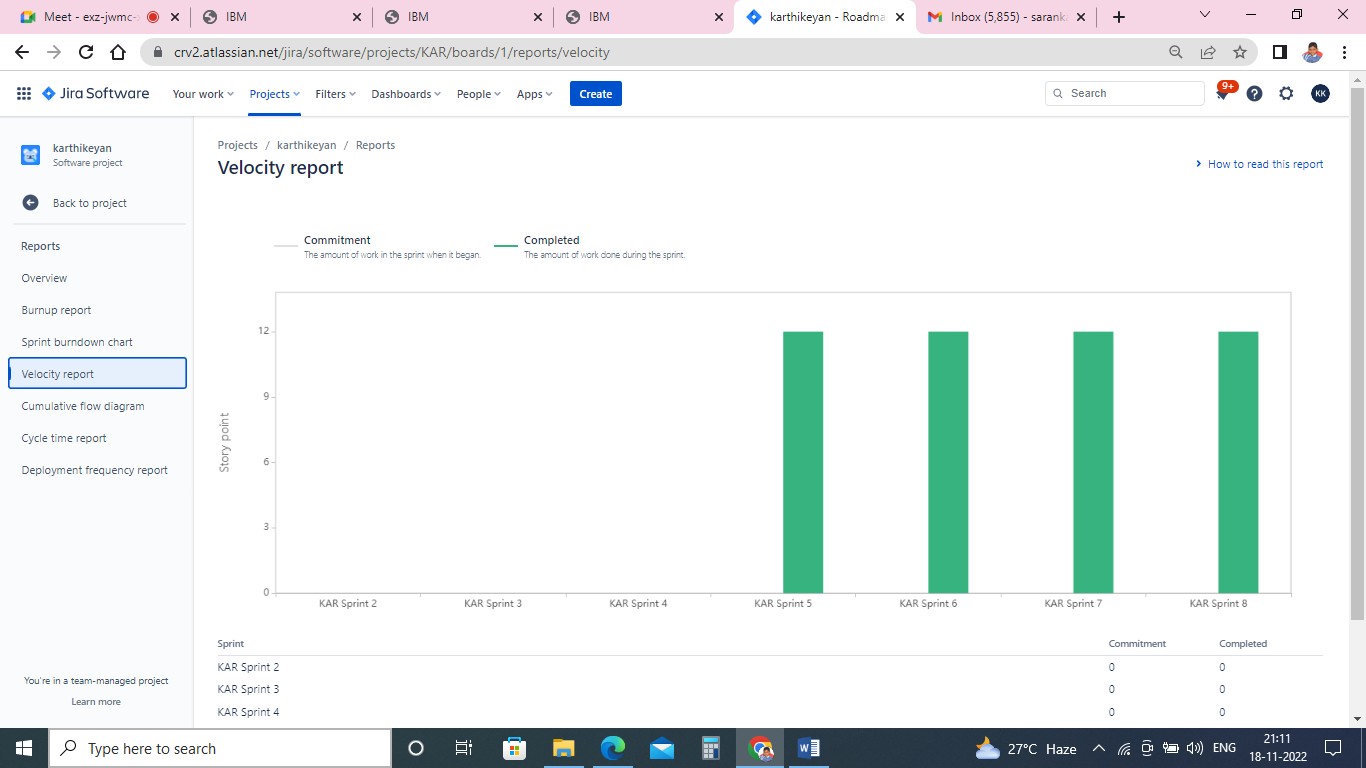
* 1. Reports from JIRA



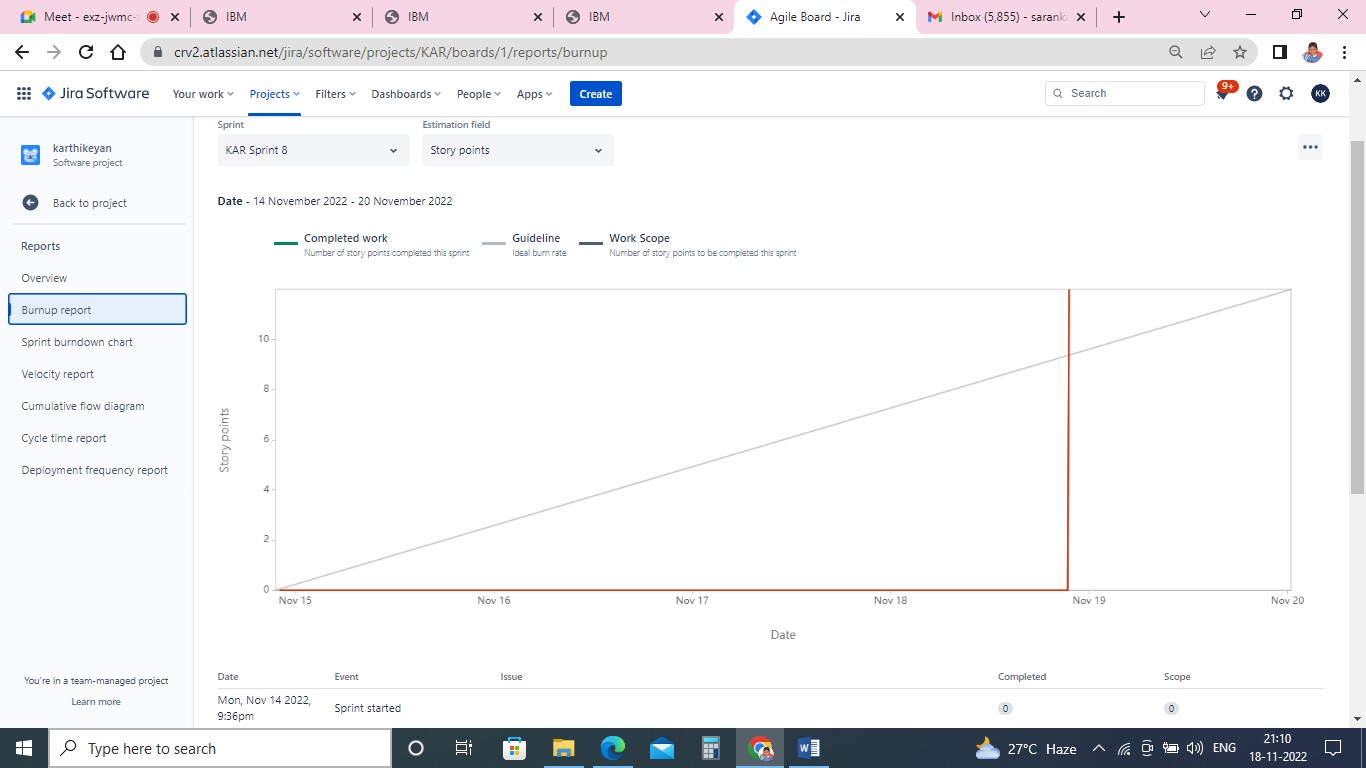
Roadmap



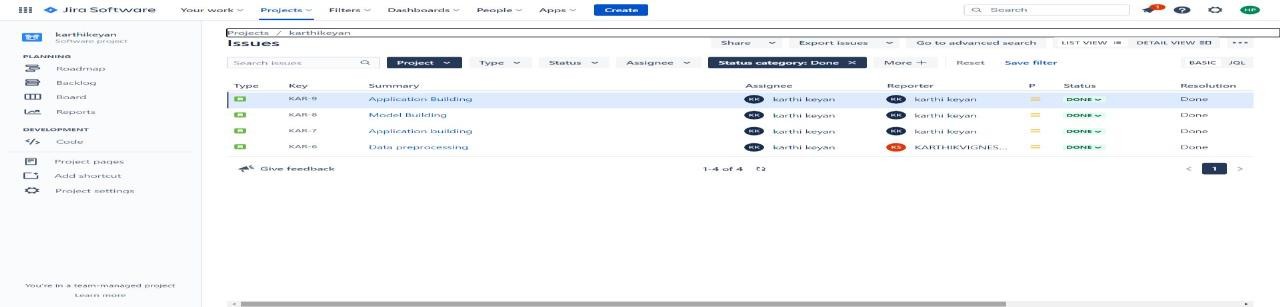
Cumulative flow diagram 11



Velocity Graph

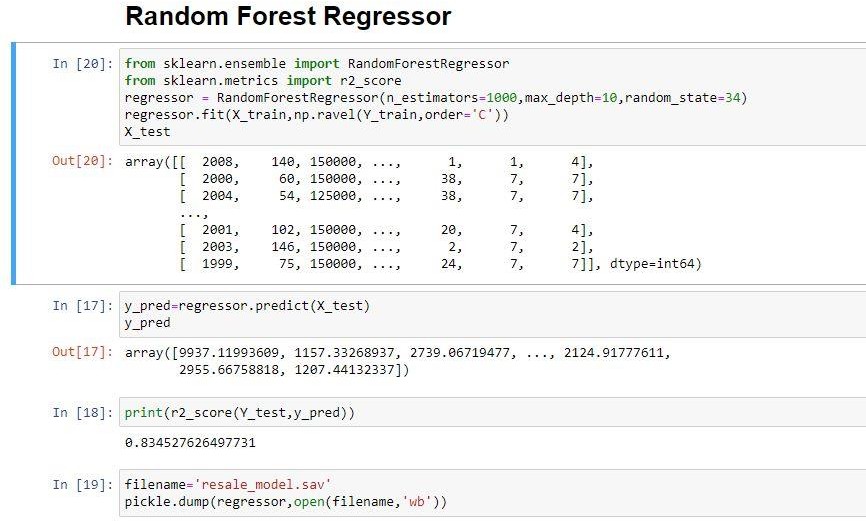


Burnup Report



Completed report 12

## Feature 1

Regression Model:

* 1. Feature 2 Accuracy:

**CHAPTER 7 CODING AND SOLUTIONING**

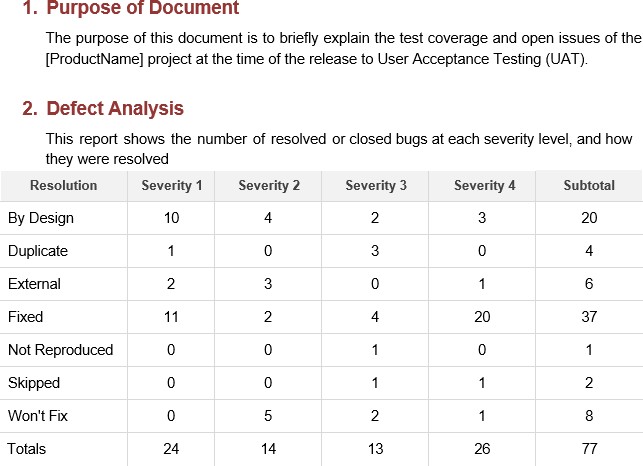


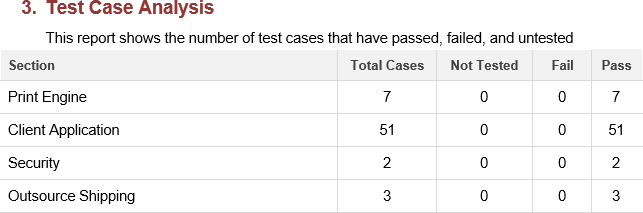
13

## Test Cases

## User Acceptance Testing

**CHAPTER 8 TESTING**





## Performance Metrics

# CHAPTER 9 RESULTS

|  |  |  |  |
| --- | --- | --- | --- |
| **S.N**  **o.** | **Parameter** | **Values** | **Screenshot** |
| 1. | Metrics | **Regression Model:**  MAE - , MSE - , RMSE -  , R2 score -  **Classification Model:** Confusion Matrix - , Accuray Score- & Classification Report –  Label Encoder |  |

15

ADVANTAGES

# CHAPTER 10

* + - The model is to be built that would give the nearest resale value of the vehicle.
    - By using these model best accuracy value will be taken as a solution for the given used car.
    - Used car price prediction is effectively used to determine the worthiness of the car by their own within few minutes by giving different parameters as input such as year, model, km driven, etc.
    - A loss function is to be optimized and mainly a weak learner can make predictions for used cars easily.
    - Using Stored data and machine learning approaches, this project proposed a scalable framework for predicting values for different type of used cars present all over India.
    - It helps users to predict the correct valuation of the car remotely with perfect valuation and without human intervention like car dealers in the process to eliminate biased valuation predicted by the dealer

## DISADVANTAGES

* + - Even though predicting the value of a used car by these model Taking a proper test drive before buying a used car will show a perfect condition of the car to predict the value by their own.
    - The main limitation of random forest is that a large number of trees can make the algorithm too slow and ineffective for real-time predictions.
    - In general, these algorithms are fast to train, but quite slow to create predictions once they are trained.
    - They are largely unstable compared to other decision predictors.

# CHAPTER 11 CONCLUSION

The increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. The proposed system will help to determine the accurate price of used car price prediction.

16

# CHAPTER 12 FUTURE SCOPE

In future this machine learning model may bind with various website which can provide real time data for price prediction. Also we may add large historical data of car price which can help to improve accuracy of the machine learning model. We can build an android app as user interface for interacting with user. For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset.

## Source Code

Model building:

**import** pandas **as** pd **import** numpy **as** np **import** matplotlib **as** plt

**from** sklearn.preprocessing **import** LabelEncoder

**import** pickle **import** os**,** types **import** pandas **as** pd

**from** botocore.client **import** Config

**import** ibm\_boto3

**def** iter (self): **return** 0

# CHAPTER 13 APPENDIX

*# @hidden\_cell*

*# The following code accesses a file in your IBM Cloud Object Storage. It includes your credentials. # You might want to remove those credentials before you share the notebook.*

cos\_client **=** ibm\_boto3**.**client(service\_name**=**'s3', ibm\_api\_key\_id**=**'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx', ibm\_auth\_endpoint**=**"https://iam.cloud.ibm.com/oidc/token", config**=**Config(signature\_version**=**'oauth'), endpoint\_url**=**'https://s3.private.us.cloud-object-storage.appdomain.cloud')

bucket **=** 'carresalevaluepredictiondeploymen-donotdelete-pr-ryosh4pvhxemjh' object\_key **=** 'autos\_preprocessed.csv'

body **=** cos\_client**.**get\_object(Bucket**=**bucket,Key**=**object\_key)['Body']

*# add missing iter method, so pandas accepts body as file-like object*

**if not** hasattr(body, " iter "): body**.** iter **=** types**.**MethodType( iter , body )

df **=** pd**.**read\_csv(body) df**.**head()

*# print(df.seller.value\_counts()) # df[df.seller!='gewerblich']*

*# # df=df.drop('seller',1)*

*# print(df.offerType.value\_counts()) # df[df.offerType!='Gesuch']*

*# df=df.drop('offerType',1) 17*

print(df**.**shape)

df**=**df[(df**.**powerPS **>**50) **&** (df**.**powerPS **<**900)]

df**=**df[(df**.**yearOfRegistration **>=**1950) **&** (df**.**yearOfRegistration **<**2017)] print(df**.**shape)

(278578, 12)

(278578, 12)

*# df.drop(['name','abtest','dateCrawled','nrOfPictures','lastSeen','postalCode','dateCreated'],axis='columns',inplace=True)*

new\_df**=**df**.**copy() new\_df**=**new\_df**.**drop\_duplicates(['price','vehicleType','yearOfRegistration','gearbox','powerPS','model','kilometer','monthOfRegistr ation','fuelType','notRepairedDamage'])

new\_df**.**gearbox**.**replace(('manuell','automatik'),('manual','automatic'),inplace**=True**) new\_df**.**fuelType**.**replace(('benzin','andere','elektro'),('petrol','others','electric'),inplace**=True**) new\_df**.**vehicleType**.**replace(('kleinwagen','cabrio','kombi','andere'),('small car','convertible','combination','others'),inplace**=True**) new\_df**.**notRepairedDamage**.**replace(('ja','nein'),('Yes','No'),inplace**=True**)

new\_df**=**new\_df[(new\_df**.**price **>=**100)**&**(new\_df**.**price **<=**150000)] new\_df['notRepairedDamage']**.**fillna(value**=**'not-declared',inplace**=True**) new\_df['fuelType']**.**fillna(value**=**'not-declared',inplace**=True**) new\_df['gearbox']**.**fillna(value**=**'not-declared',inplace**=True**)

*# new\_df['vehicleType'].fillna(value='not-declared',inplace=True)*

new\_df['model']**.**fillna(value**=**'not-declared',inplace**=True**)

*# new\_df.to\_csv("autos\_preprocessed.csv")*

**from** sklearn.preprocessing **import** LabelEncoder labels**=**['gearbox','notRepairedDamage','model','brand','fuelType','vehicleType'] mapper**=**{}

**for** i **in** labels:

mapper[i]**=**LabelEncoder() mapper[i]**.**fit(new\_df[i]) tr**=**mapper[i]**.**transform(new\_df[i]) np**.**save(str('classes'**+**i**+**'.npy'),mapper[i]**.**classes\_) print(i,":",mapper[i])

new\_df**.**loc[:,i**+**'\_labels']**=**pd**.**Series(tr,index**=**new\_df**.**index)

labeled **=** new\_df[['price','yearOfRegistration','powerPS','kilometer','monthOfRegistration']**+**[x**+**"\_labels" **for** x **in** labels]] print(labeled**.**columns)

gearbox : LabelEncoder() notRepairedDamage : LabelEncoder() model : LabelEncoder()

brand : LabelEncoder() fuelType : LabelEncoder() vehicleType : LabelEncoder()

Index(['price', 'yearOfRegistration', 'powerPS', 'kilometer', 'monthOfRegistration', 'gearbox\_labels', 'notRepairedDamage\_labels', 'model\_labels', 'brand\_labels', 'fuelType\_labels',

'vehicleType\_labels'], dtype='object')

Y**=**labeled**.**iloc[:,0]**.**values X**=**labeled**.**iloc[:,1:]**.**values

Y**=**Y**.**reshape(**-**1,1)

**from** sklearn.model\_selection **import** cross\_val\_score,train\_test\_split

18

X\_train,X\_test,Y\_train,Y\_test**=** train\_test\_split(X,Y,test\_size**=**0.3,random\_state**=**3)

**from** sklearn.ensemble **import** RandomForestRegressor

**from** sklearn.metrics **import** r2\_score

regressor **=** RandomForestRegressor(n\_estimators**=**1000,max\_depth**=**10,random\_state**=**34) regressor**.**fit(X\_train,np**.**ravel(Y\_train,order**=**'C'))

RandomForestRegressor(max\_depth=10, n\_estimators=1000, random\_state=34) y\_pred**=**regressor**.**predict(X\_test)

print(r2\_score(Y\_test,y\_pred))

0.834527626497731

filename**=**'resale\_model.sav' pickle**.**dump(regressor,open(filename,'wb'))

## GitHub link: [Click here](https://github.com/IBM-EPBL/IBM-Project-23383-1659881456)

Project Demo Link: [Click here](https://github.com/IBM-EPBL/IBM-Project-23383-1659881456/tree/main/Final%20Deliverables/Demonstration%20video)

19