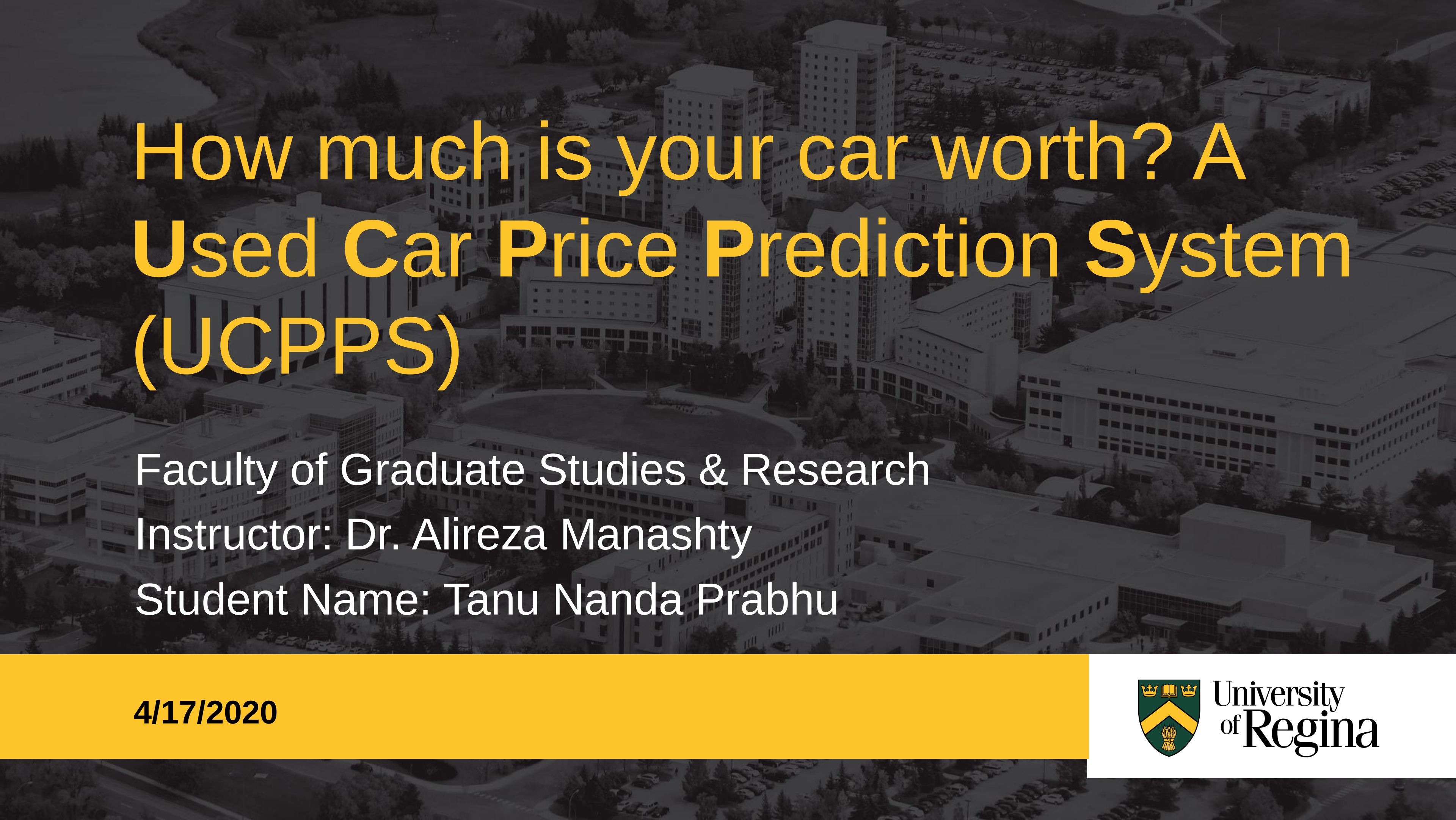


How much is your car worth? A **U**sed **C**ar **P**rice **P**rediction **S**ystem (UCPPS)

Faculty of Graduate Studies & Research

Instructor: Dr. Alireza Manashty

Student Name: Tanu Nanda Prabhu

4/17/2020

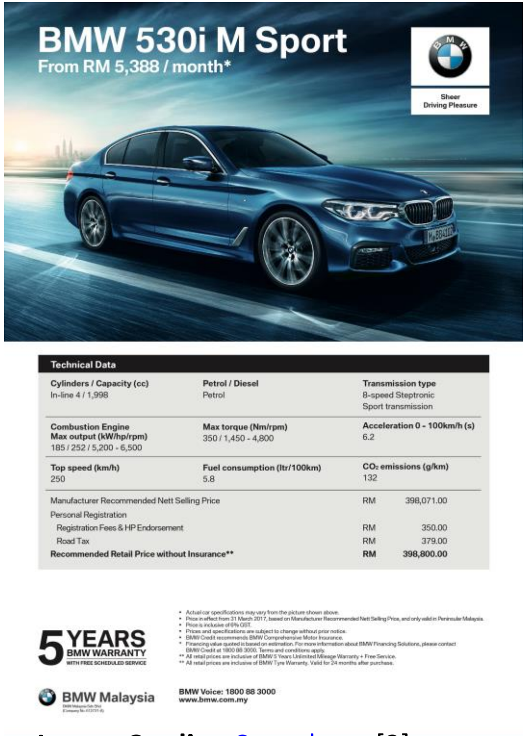
University of Regina

Table of Contents

- Introduction
- Problem Statement
- Users
- Situation and Project Goals
- Executive Summary
- Approach
- Data and Model Description
- Tools
- Team Role and Timeline
- Application and Recommendation
- References

Introduction

- Vehicle value forecast is a significant errand particularly when the vehicle is used.
- The value of the car depends on several factors:
    - Make (brand of the car)
    - Power
    - Number of kilometers it has been run
    - Year of registration, and many more
- Better the features higher the price



BMW 530i M Sport
From RM 5,388 / month*

Sheer
Driving Pleasure

Technical Data

Cylinders / Capacity (cc) In-line 4 / 1,998	**Petrol / Diesel** Petrol	**Transmission type** 8-speed Steptronic Sport transmission
Combustion Engine **Max output (kW/hp/rpm)** 185 / 252 / 5,200 - 6,500	**Max torque (Nm/rpm)** 350 / 1,450 - 4,800	**Acceleration 0 - 100km/h (s)** 6.2
Top speed (km/h) 250	**Fuel consumption (ltr/100km)** 5.8	**CO_2 emissions (g/km)** 132

Manufacturer Recommended Nett Selling Price	RM	398,071.00
Personal Registration		
Registration Fees & HP Endorsement	RM	350.00
Road Tax	RM	379.00
Recommended Retail Price without Insurance**	**RM**	**398,800.00**

5 YEARS BMW WARRANTY
WITH FREE SCHEDULED SERVICE

- Actual car specifications may vary from the picture shown above.
- Price in effect from 31 March 2017, based on Manufacturer Recommended Nett Selling Price, and only valid in Peninsular Malaysia.
- Price is inclusive of 6% GST.
- Prices and specifications are subject to change without prior notice.
- BMW Credit recommends BMW Comprehensive Motor Insurance.

* Financing value quoted is based on estimation. For more information about BMW Financing Solutions, please contact BMW Credit at 1800 88 3000. Terms and conditions apply.
** All retail prices are inclusive of BMW 5 Years Unlimited Mileage Warranty + Free Service.
** All retail prices are inclusive of BMW Tyre Warranty. Valid for 24 months after purchase.

BMW Malaysia
BMW Malaysia Sdn Bhd
(Company No. [illegible])

BMW Voice: 1800 88 3000
www.bmw.com.my

Image Credits: Specsheet [2]

Problem Statement

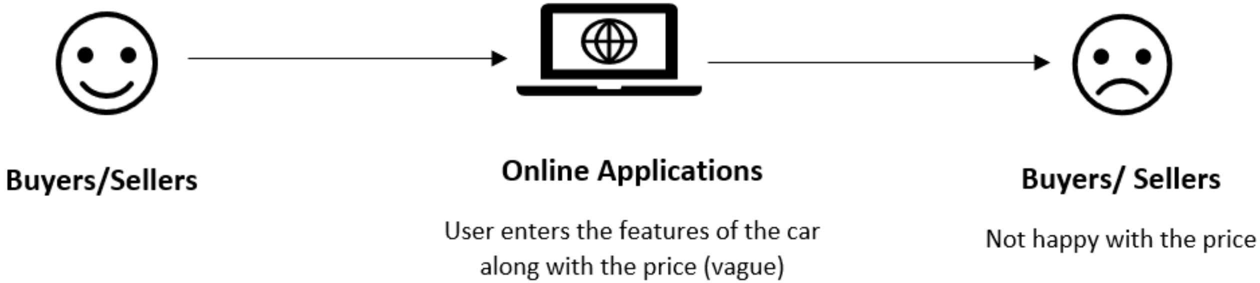
•Used Car Prices are important reflection of the economy and they greatly interest both buyers and sellers.

•A prediction model that estimates resale price based on car’s attributes or features is much more needed today.

•My analysis aims to determine which features of the car that may have the strongest statistical correlation with the price of the car.

Problem Statement

- Current Situation



Buyers/Sellers

Online Applications

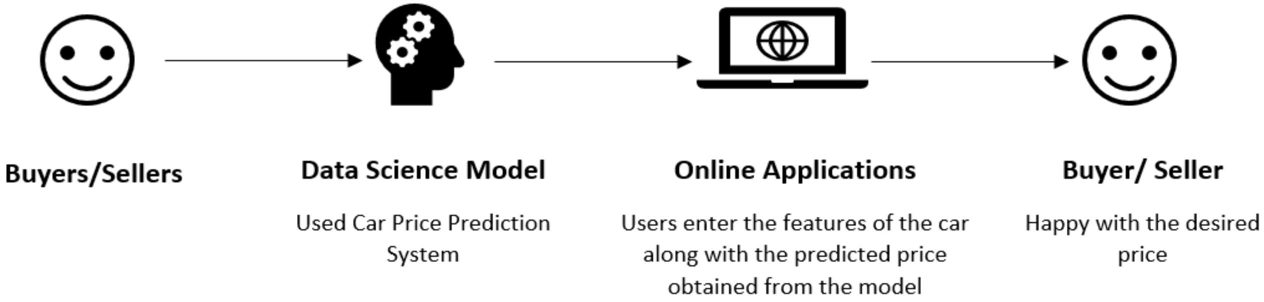
User enters the features of the car along with the price (vague)

Buyers/ Sellers

Not happy with the price

Problem Statement

- Desired Situation



Buyers/Sellers

Data Science Model

Used Car Price Prediction System

Online Applications

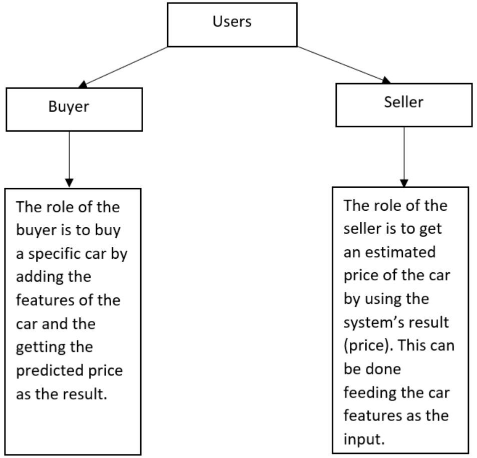
Users enter the features of the car along with the predicted price obtained from the model

Buyer/ Seller

Happy with the desired price

Users

- Two Types of Users
  - Buyer
  - Seller
- Both of the users don’t have to worry about paying excess or end getting less paid.
- Canada Used Car Dealer Retail Sales is at a level of 1.056B CAD for Nov 2019 [3]

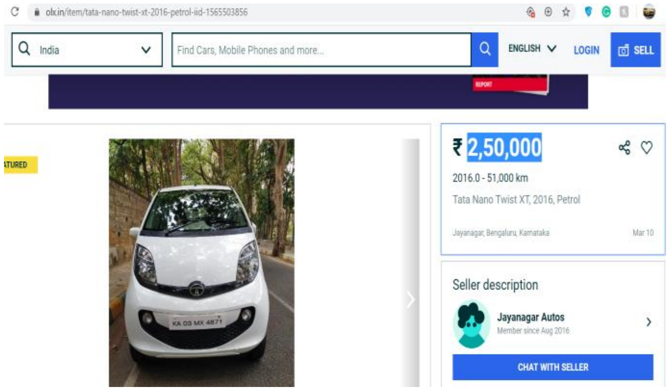


The role of the buyer is to buy a specific car by adding the features of the car and the getting the predicted price as the result.

The role of the seller is to get an estimated price of the car by using the system’s result (price). This can be done feeding the car features as the input.

Situation and Project Goals

- **Situation**
  - Currently, the user feeds the features of their car as input in online applications.
  - For price, they have to guess, or just give high price values.
  - No mechanisms to predict the price of the car given the features. So nobody will buy the car because of high price.
  - Waste of time


olx.in/item/tata-nano-twist-xt-2016-petrol-iid-1565503856
India
Find Cars, Mobile Phones and more...
ENGLISH
LOGIN
SELL
REPORT
ATURED
KA 03 MX 4871
CHAT WITH SELLER

₹ 2,50,000

2016.0 - 51,000 km

Tata Nano Twist XT, 2016, Petrol

Jayanagar, Bengaluru, Karnataka Mar 10

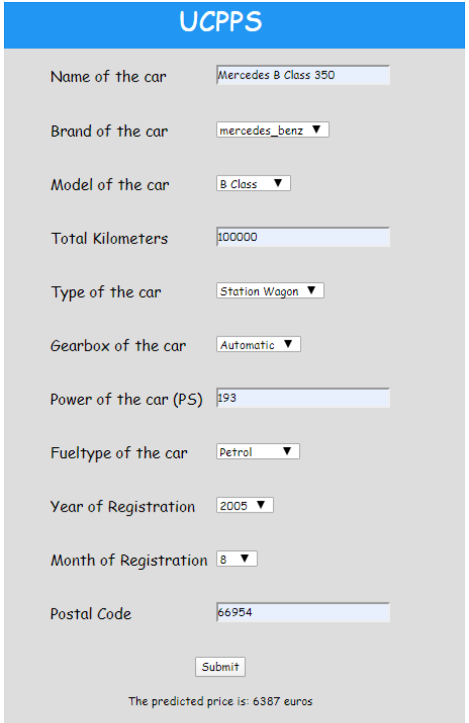
Seller description

Jayanagar Autos
Member since Aug 2016

Image: https://www.olx.in/item/tata-nano-twist-xt-2016-petrol-iid-1565503856

Situation and Project Goals

- **Project Goals**
  - – Predicting the price based on the given features.
  - – Users can get a estimated price, no need of guessing or setting a random price.
  - – More chances that the car is purchased because of estimated price by the system
  - – No waste of time.



UCPPS

Name of the car	Mercedes B Class 350
Brand of the car	mercedes_benz ▼
Model of the car	B Class ▼
Total Kilometers	100000
Type of the car	Station Wagon ▼
Gearbox of the car	Automatic ▼
Power of the car (PS)	193
Fueltype of the car	Petrol ▼
Year of Registration	2005 ▼
Month of Registration	8 ▼
Postal Code	66954

The predicted price is: 6387 euros

Executive Summary

Used Car Price Prediction Website predicts price with 86% accuracy

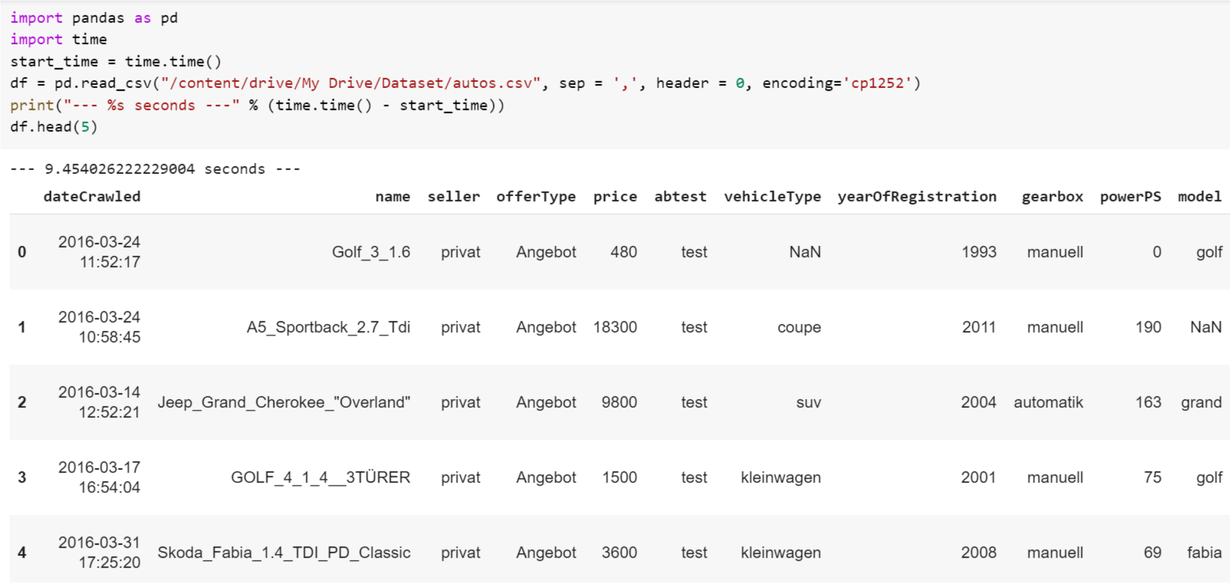
- The key features were recognized from the new website
  - Users can get the estimated price of their car faster than ever.
  - Anyone who is interested to know how much their car is worth for can use it with ease
  - Saves a lot of money to both buyers and seller with the help of the system
- We can make reasonably accurate predictions with limited data
  - Brand, Model, Kilometer and Year of Registration are most influential factors and predicts outcomes accurately 86% of the time.
  - The best prediction model was chosen to implement (Random Forest Regression)
- Model can run online with 2-3 seconds
  - Model performance is finely tuned by filtering various parameters
  - Initially, the model take few more seconds to load the website, but further testing took less time.

Approach

- Drafting the problem statement
- Taking the existing used car data set from the internet.
- Followed all the data analytic life cycle
- Developed a predictive model to predict the price of the used car given the features.
    - Identified most influential factors (dependent features)
    - Model is very accurate in predicting the price
- Developed a website to simulate model performance

Data

- The data set was chosen from data.world, which was originally scraped from e-bay [4]

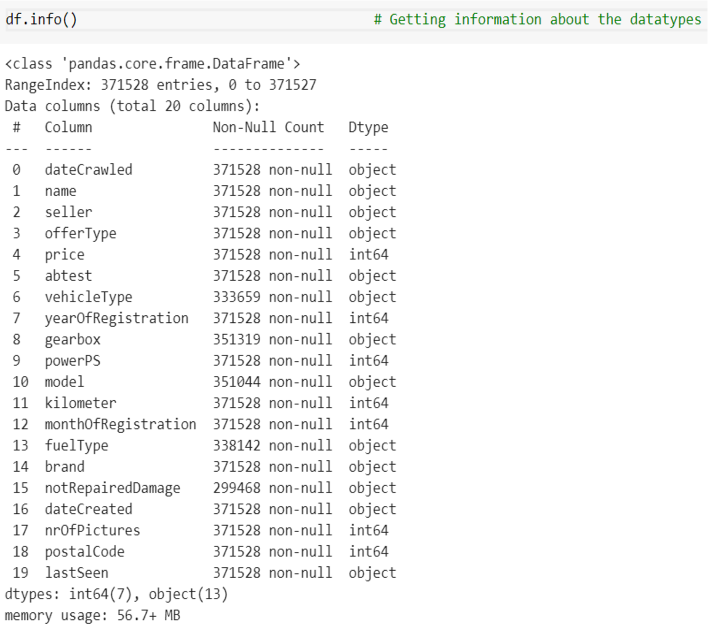


```
import pandas as pd
import time
start_time = time.time()
df = pd.read_csv("/content/drive/My Drive/Dataset/autos.csv", sep = ',', header = 0, encoding='cp1252')
print("--- %s seconds ---" % (time.time() - start_time))
df.head(5)
```

```
--- 9.454026222229004 seconds ---
```

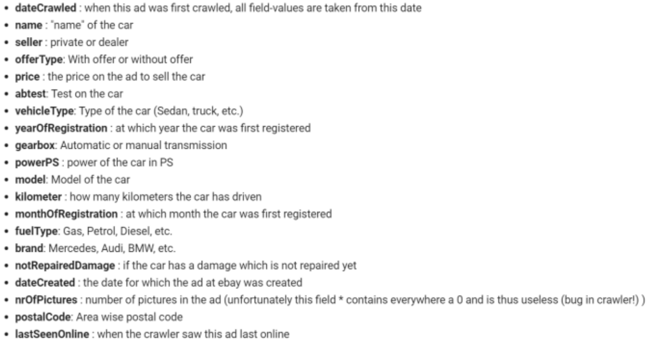
	dateCrawled	name	seller	offerType	price	abtest	vehicleType	yearOfRegistration	gearbox	powerPS	model
0	2016-03-24 11:52:17	Golf_3_1.6	privat	Angebot	480	test	NaN	1993	manuell	0	golf
1	2016-03-24 10:58:45	A5_Sportback_2.7_Tdi	privat	Angebot	18300	test	coupe	2011	manuell	190	NaN
2	2016-03-14 12:52:21	Jeep_Grand_Cherokee_"Overland"	privat	Angebot	9800	test	suv	2004	automatik	163	grand
3	2016-03-17 16:54:04	GOLF_4_1_4__3TÜRER	privat	Angebot	1500	test	kleinwagen	2001	manuell	75	golf
4	2016-03-31 17:25:20	Skoda_Fabia_1.4_TDI_PD_Classic	privat	Angebot	3600	test	kleinwagen	2008	manuell	69	fabia

Data Statistics

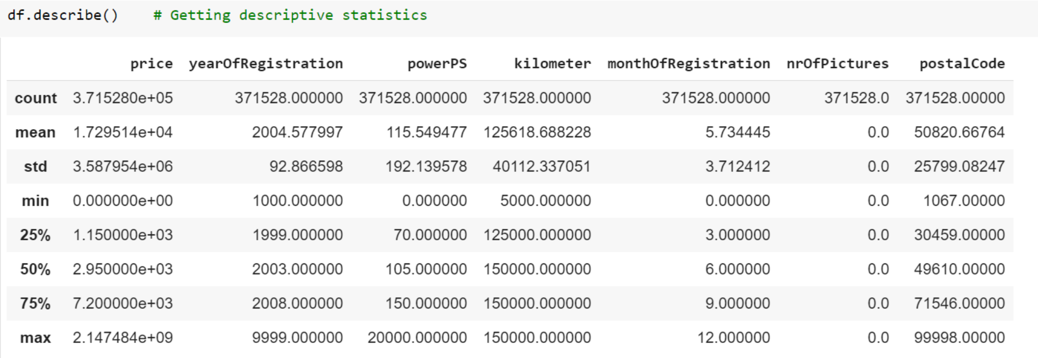


```
df.info()                                   # Getting information about the datatypes
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 371528 entries, 0 to 371527
Data columns (total 20 columns):
 #   Column               Non-Null Count   Dtype 
---  ------               --------------   ----- 
 0   dateCrawled          371528 non-null  object
 1   name                 371528 non-null  object
 2   seller               371528 non-null  object
 3   offerType            371528 non-null  object
 4   price                371528 non-null  int64 
 5   abtest               371528 non-null  object
 6   vehicleType          333659 non-null  object
 7   yearOfRegistration   371528 non-null  int64 
 8   gearbox              351319 non-null  object
 9   powerPS              371528 non-null  int64 
 10  model                351044 non-null  object
 11  kilometer            371528 non-null  int64 
 12  monthOfRegistration  371528 non-null  int64 
 13  fuelType             338142 non-null  object
 14  brand                371528 non-null  object
 15  notRepairedDamage    299468 non-null  object
 16  dateCreated          371528 non-null  object
 17  nrOfPictures         371528 non-null  int64 
 18  postalCode           371528 non-null  int64 
 19  lastSeen             371528 non-null  object
dtypes: int64(7), object(13)
memory usage: 56.7+ MB
```



- **dateCrawled** : when this ad was first crawled, all field-values are taken from this date
- **name** : "name" of the car
- **seller** : private or dealer
- **offerType**: With offer or without offer
- **price** : the price on the ad to sell the car
- **abtest**: Test on the car
- **vehicleType**: Type of the car (Sedan, truck, etc.)
- **yearOfRegistration** : at which year the car was first registered
- **gearbox**: Automatic or manual transmission
- **powerPS** : power of the car in PS
- **model**: Model of the car
- **kilometer** : how many kilometers the car has driven
- **monthOfRegistration** : at which month the car was first registered
- **fuelType**: Gas, Petrol, Diesel, etc.
- **brand**: Mercedes, Audi, BMW, etc.
- **notRepairedDamage** : if the car has a damage which is not repaired yet
- **dateCreated** : the date for which the ad at ebay was created
- **nrOfPictures** : number of pictures in the ad (unfortunately this field * contains everywhere a 0 and is thus useless (bug in crawler!))
- **postalCode**: Area wise postal code
- **lastSeenOnline** : when the crawler saw this ad last online



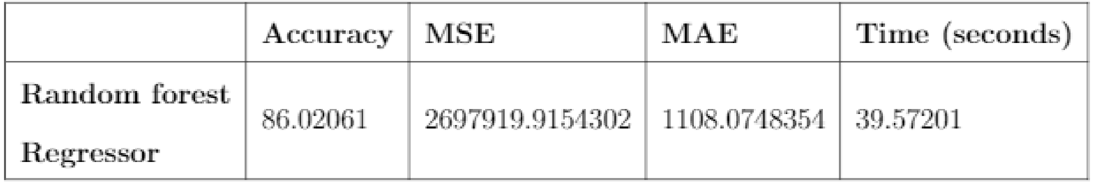
```
df.describe()    # Getting descriptive statistics
```

	price	yearOfRegistration	powerPS	kilometer	monthOfRegistration	nrOfPictures	postalCode
count	3.715280e+05	371528.000000	371528.000000	371528.000000	371528.000000	371528.0	371528.00000
mean	1.729514e+04	2004.577997	115.549477	125618.688228	5.734445	0.0	50820.66764
std	3.587954e+06	92.866598	192.139578	40112.337051	3.712412	0.0	25799.08247
min	0.000000e+00	1000.000000	0.000000	5000.000000	0.000000	0.0	1067.00000
25%	1.150000e+03	1999.000000	70.000000	125000.000000	3.000000	0.0	30459.00000
50%	2.950000e+03	2003.000000	105.000000	150000.000000	6.000000	0.0	49610.00000
75%	7.200000e+03	2008.000000	150.000000	150000.000000	9.000000	0.0	71546.00000
max	2.147484e+09	9999.000000	20000.000000	150000.000000	12.000000	0.0	99998.00000

Model Building

- Linear Regression
- Support Vector Machine Regression
- Random Forest Regression
- Decision Tree Regression

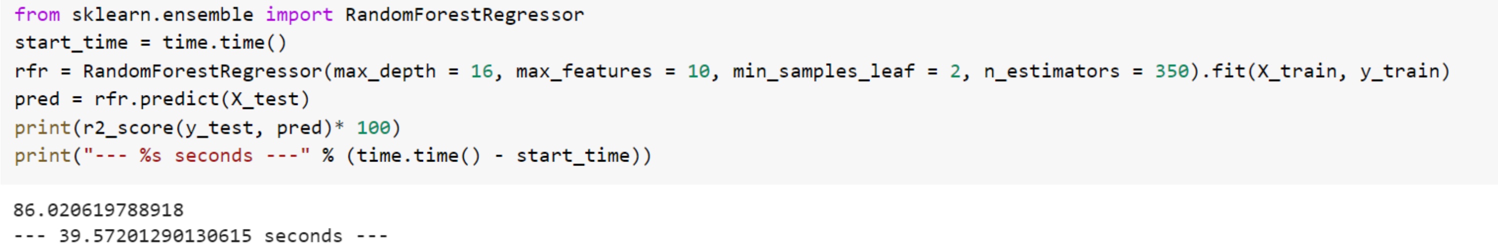
- Winner was – **Random Forest Regression**



	Accuracy	MSE	MAE	Time (seconds)
Random forest Regressor	86.02061	2697919.9154302	1108.0748354	39.57201

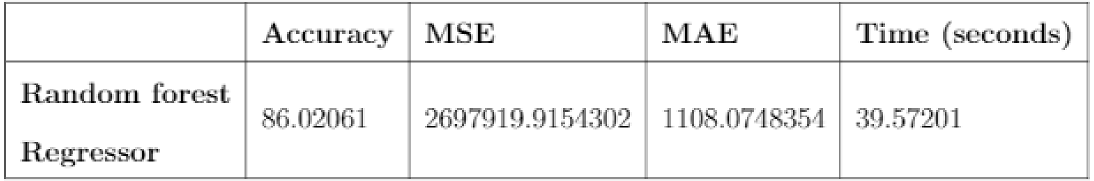
Model Building

- Model Performance Assessment
  - Final Model with the best parameters



```python
from sklearn.ensemble import RandomForestRegressor
start_time = time.time()
rfr = RandomForestRegressor(max_depth = 16, max_features = 10, min_samples_leaf = 2, n_estimators = 350).fit(X_train, y_train)
pred = rfr.predict(X_test)
print(r2_score(y_test, pred)* 100)
print("--- %s seconds ---" % (time.time() - start_time))
```

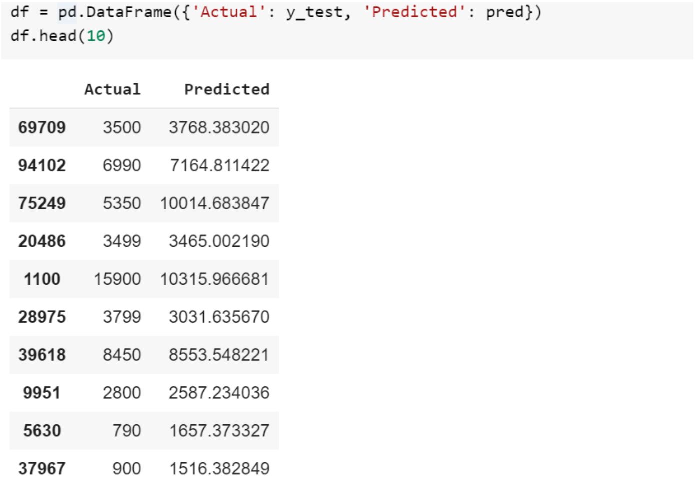
```
86.020619788918
--- 39.57201290130615 seconds ---
```



	Accuracy	MSE	MAE	Time (seconds)
Random forest Regressor	86.02061	2697919.9154302	1108.0748354	39.57201

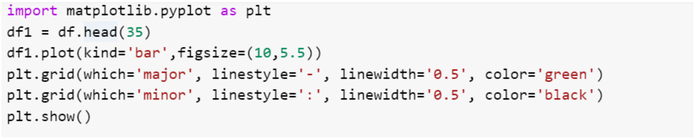
Model Results

- Comparing the actual values vs predicted values
- Training accuracy was 99% and testing accuracy was 86%

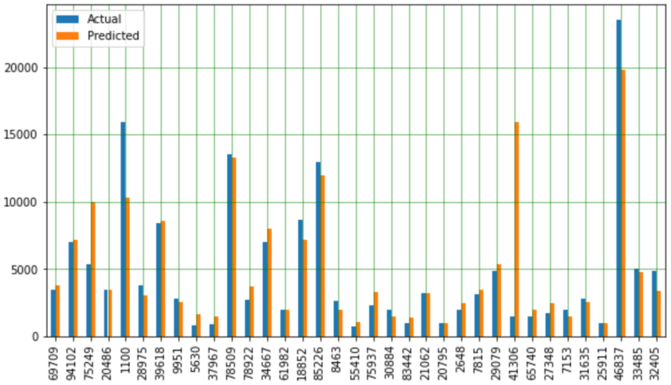


```
df = pd.DataFrame({'Actual': y_test, 'Predicted': pred})
df.head(10)
```

	Actual	Predicted
69709	3500	3768.383020
94102	6990	7164.811422
75249	5350	10014.683847
20486	3499	3465.002190
1100	15900	10315.966681
28975	3799	3031.635670
39618	8450	8553.548221
9951	2800	2587.234036
5630	790	1657.373327
37967	900	1516.382849

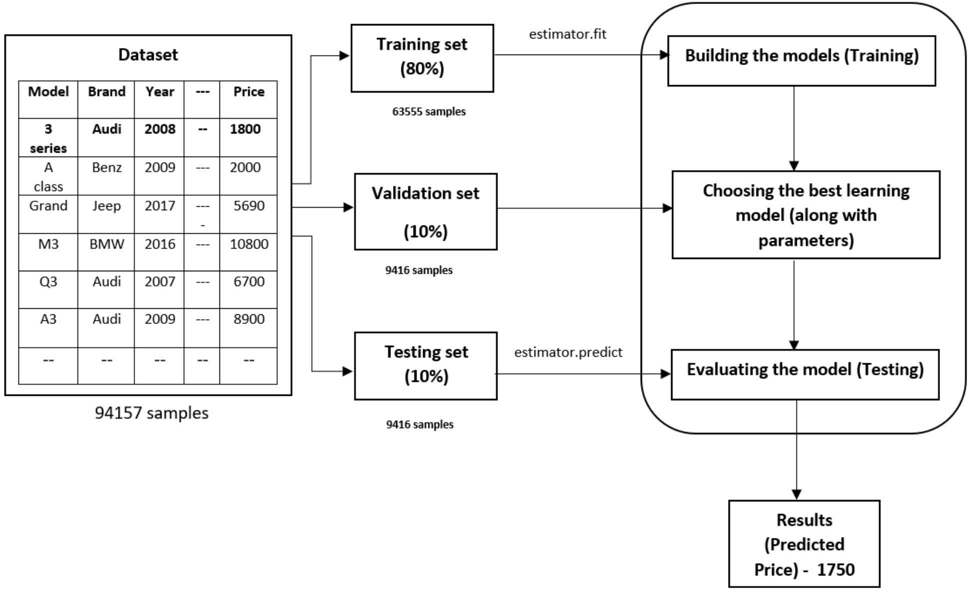


```
import matplotlib.pyplot as plt
df1 = df.head(35)
df1.plot(kind='bar',figsize=(10,5.5))
plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
plt.show()
```



Solution Overview

- Category of data falls under supervised machine learning: Regression Estimator



Dataset

Model	Brand	Year	---	Price
3 series	**Audi**	**2008**	**--**	**1800**
A class	Benz	2009	---	2000
Grand	Jeep	2017	----	5690
M3	BMW	2016	---	10800
Q3	Audi	2007	---	6700
A3	Audi	2009	---	8900
--	--	--	--	--

94157 samples

Training set (80%)
63555 samples
Validation set (10%)
9416 samples
Testing set (10%)
9416 samples
estimator.fit
estimator.predict
Building the models (Training)
Choosing the best learning model (along with parameters)
Evaluating the model (Testing)
Results (Predicted Price) - 1750

Model Description

- **Overview of Methodology**
  - Predict the price for each car based on its features
  - Divided the data into three sets (Train, Test and Validation)

- **Best model**–Random Forest Regression
- **Dependent Variables**–Price
- **Independent Variables** – Features of the car such as brand, model, year and month of registration, etc.

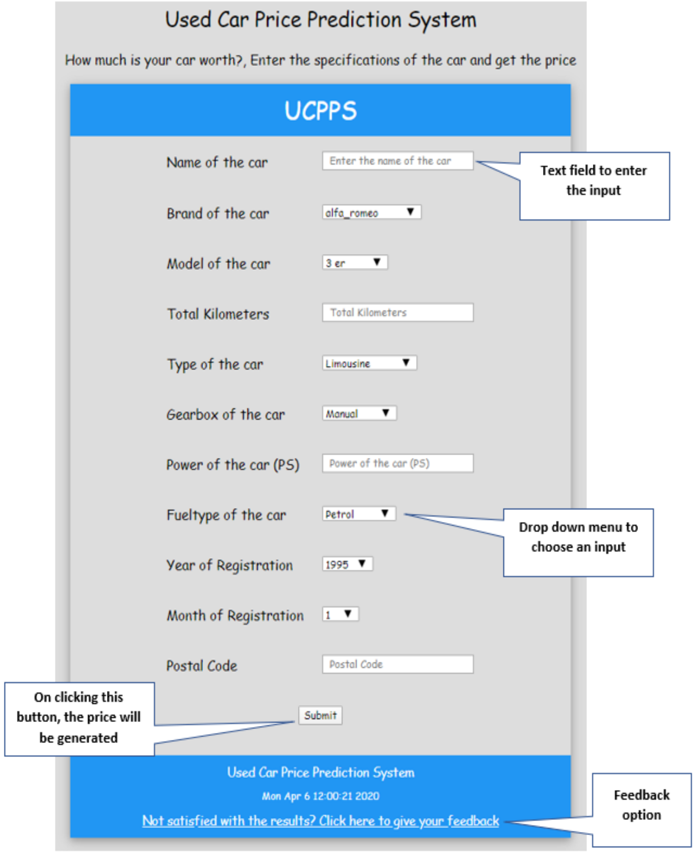
- **Scope and sample**
  - 371, 528 instances and 20 features.
  - Training sample 90% of data
  - Testing and validation sample 5% and 5% data

- **The model developed has reasonable predictive power for the data set provided**
  - Brand, Model, Kilometers, Year and month of Registration, Horsepower and many more are most influential factors and predict the outcome (price) 86% of the time
  - Model shown marginal improvement on new data when tested

UCPPS Web Application

- Web application was developed using Django Web framework on Heroku cloud.

- Link: https://ucpps.herokuapp.com/ucpps_app/home.html



Used Car Price Prediction System

How much is your car worth?, Enter the specifications of the car and get the price

UCPPS

Name of the car	Enter the name of the car
Brand of the car	alfa_romeo ▼
Model of the car	3 er ▼
Total Kilometers	Total Kilometers
Type of the car	Limousine ▼
Gearbox of the car	Manual ▼
Power of the car (PS)	Power of the car (PS)
Fueltype of the car	Petrol ▼
Year of Registration	1995 ▼
Month of Registration	1 ▼
Postal Code	Postal Code

Submit

Used Car Price Prediction System

Mon Apr 6 12:00:21 2020

Not satisfied with the results? Click here to give your feedback

Text field to enter the input

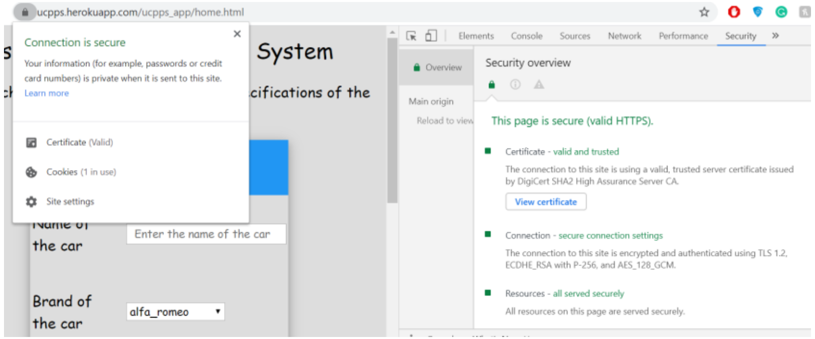
Drop down menu to choose an input

On clicking this button, the price will be generated

Feedback option

UCPPS Application

- Features of the application



ucpps.herokuapp.com/ucpps_app/home.html

Connection is secure

Your information (for example, passwords or credit card numbers) is private when it is sent to this site.

Certificate (Valid)

Cookies (1 in use)

Site settings

Security overview

This page is secure (valid HTTPS).

Certificate - valid and trusted

The connection to this site is using a valid, trusted server certificate issued by DigiCert SHA2 High Assurance Server CA.

Connection - secure connection settings

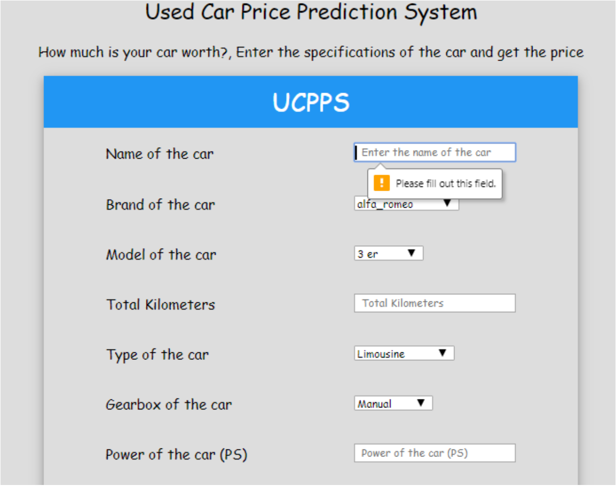
The connection to this site is encrypted and authenticated using TLS 1.2, ECDHE_RSA with P-256, and AES_128_GCM.

Resources - all served securely

All resources on this page are served securely.

Name of the car	Enter the name of the car
Brand of the car	alfa_romeo ▾

Connection link is secure



Used Car Price Prediction System

How much is your car worth?, Enter the specifications of the car and get the price

UCPPS

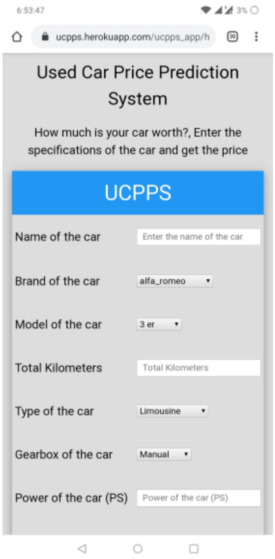
Name of the car	Enter the name of the car
Brand of the car	alfa_romeo ▼
Model of the car	3 er ▼
Total Kilometers	Total Kilometers
Type of the car	Limousine ▼
Gearbox of the car	Manual ▼
Power of the car (PS)	Power of the car (PS)

Please fill out this field.

Application is roboust

UCPPS Application

- Features of the application



6:53:47 3%

ucpps.herokuapp.com/ucpps_app/h

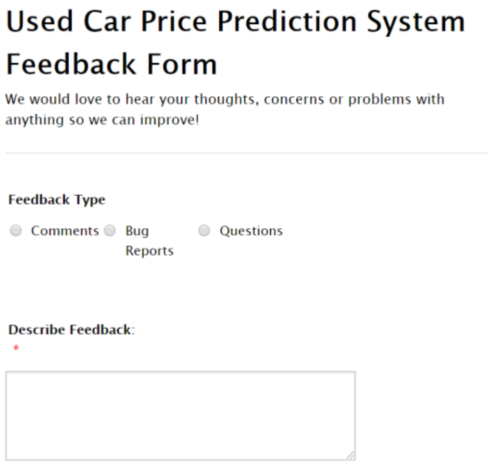
Used Car Price Prediction System

How much is your car worth?, Enter the specifications of the car and get the price

UCPPS

Name of the car	Enter the name of the car
Brand of the car	alfa_romeo ▾
Model of the car	3 er ▾
Total Kilometers	Total Kilometers
Type of the car	Limousine ▾
Gearbox of the car	Manual ▾
Power of the car (PS)	Power of the car (PS)

Responsive (Mobile Screen)



Used Car Price Prediction System Feedback Form

We would love to hear your thoughts, concerns or problems with anything so we can improve!

Feedback Type

○ Comments ○ Bug Reports ○ Questions

Describe Feedback: *

Option to give feedback

Recommendations

- Recommend this model to online applications such as CarDheko, Quikr, Olx, etc.–test and learn from model outputs.
  - Further Tuning the model can enable fast predictions and minimize the frustration by unexpected delayed predictions
  - Outputs can be further enhanced to give users advice on how to increase likelihood of successful price predictions.
- More training of the model with new dataset across the world. The more the model is trained the better its prediction would be.
- Improvements needed to the model
  - Fewer data–The more the data the better the model
  - Could have used an ensemble stacking or bagging approach to improve the accuracy
  - Could have used classification to classify different car brands according to models and then apply regression.

Recommendations

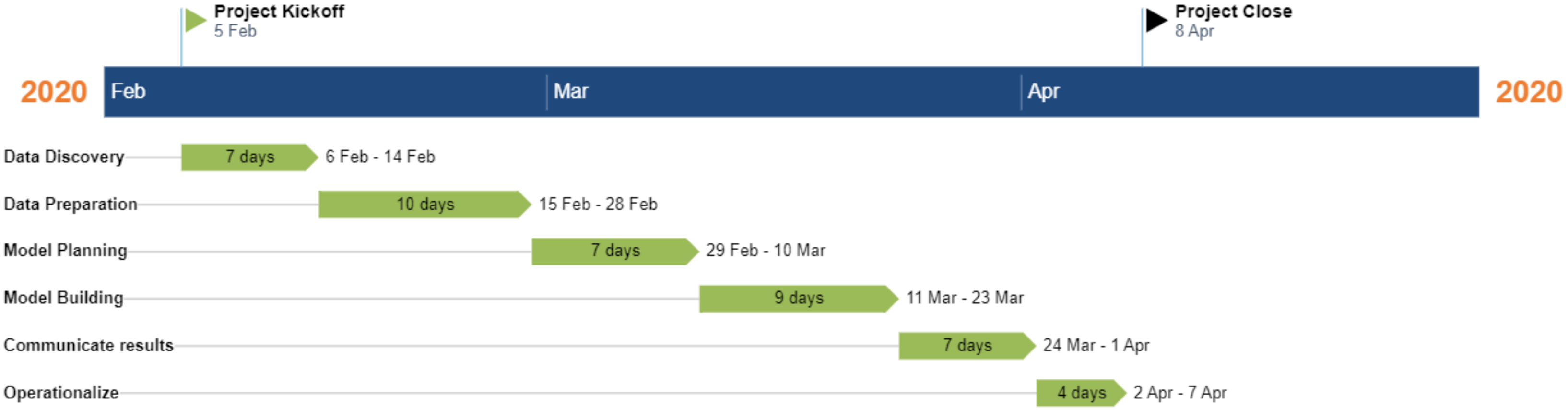
- Improvements to the website
  - Create a real time dashboard for analysts
  - To be alerted when the price is predicted along with the features (like email message)
  - Profile option could be added in the future.

- Loading time of the website can be increased by tuning the model performance.

Tools

- Google Colab
- Anaconda Jupyter Notebook
- Python Programming (Obviously)
- Python Libraries (Pandas, NumPy, Matplotlib, scikit learn)
- Django Web framework (HTML, CSS, Bootstrap)
- GitHub
- Visual Studio Code
- Heroku

Timeline



Project Kickoff
5 Feb

Project Close
8 Apr

2020 Feb Mar Apr 2020

Data Discovery 7 days 6 Feb - 14 Feb

Data Preparation 10 days 15 Feb - 28 Feb

Model Planning 7 days 29 Feb - 10 Mar

Model Building 9 days 11 Mar - 23 Mar

Communicate results 7 days 24 Mar - 1 Apr

Operationalize 4 days 2 Apr - 7 Apr

Team Roles

- Data Collection, data understanding
- Model Design, model evaluation
- Code Documentation
- Deployment and building a functional website.

Image Credits: Medium

Outcomes

•Getting a price estimation of your used car

•Not getting paid less or sold less by using this application

•Technique can also be implemented on Bikes, Trucks and other type of vehicles

•Maintenance is easy, application can be enhanced by adding car models and brands and add more extended features.

References

[1] Manashty, D. (2020). *Data Science Fundamentals - Chapter 1*. Presentation, University of Regina, Canada.

[2] *The all-new BMW 5-series (G30) launched – All You Need to Know*. (2020). [Image]. Retrieved 8 April 2020, from http://kensomuse.com/blog/2017/03/30/new-bmw-5-series-g30-launched-need-know/

[3] Canada Used Car Dealers Retail Sales. Retrieved 5 February 2020, from https://ycharts.com/indicators/canada_used_car_dealers_retail_sales

[4] Leka, O. Used Cars Data - dataset by data-society. Retrieved 8 April 2020, from https://data.world/data-society/used-cars-data

Thank You

University of Regina