

MACHINE LEARNING PIPELINE ON CLOUD

PREDICTIVE ANALYSIS

PROJECT GUIDE

Gagan Deep Singh

Assistant Professor

Department of Computer Science

University of Petroleum and Energy Studies

SUBMITTED TO

Dr. Susheela Dhaiya

Assistant Professor

Department of Computer Science

University of Petroleum and Energy Studies

TEAM MEMBERS

Name	Roll Number	SAP ID
Aditya Kumar	04	500066319
Anuj Verma	09	500069910
Mayank Joshi	25	500070105
Dhananjai Kalra	14	500070904

PROBLEM STATEMENT

- Develop and Deploy a Machine Learning Pipeline on Cloud and make it accessible through browser

GOAL

- Develop a Machine Learning Pipeline
- Create a Web Application and UI for the Pipeline
- Manage the Pipeline through VCS
- Deploy the Pipeline on Cloud Platform and make it accessible through browser

The Main Barrier to Delivering Business Value Is Lack of Successful Productizing Projects



Source: Gartner Inc.

WORKFLOW OF THE PROJECT

- Acquiring the Dataset
- Importing the Dataset into Jupyter Notebook and performing EDA, Cleaning and Basic Visualization to understand the data
- Get insights to the data and develop a suitable Machine Learning Model
- Training the ML Model and Testing for accuracy
- Writing the trained ML Model to a Pickle (.pkl) file
- Developing a WebApp and Single page website for the project
- Committing the whole project on GitHub and integrate it with Heroku
- Deploy the WebApp

TOOLS USED

Python Programming Language

Machine Learning Algorithm

Flask Framework

Jupyter Notebook

Heroku Cloud

GitHub (Version Control System)

Tableau (Data Visualization)

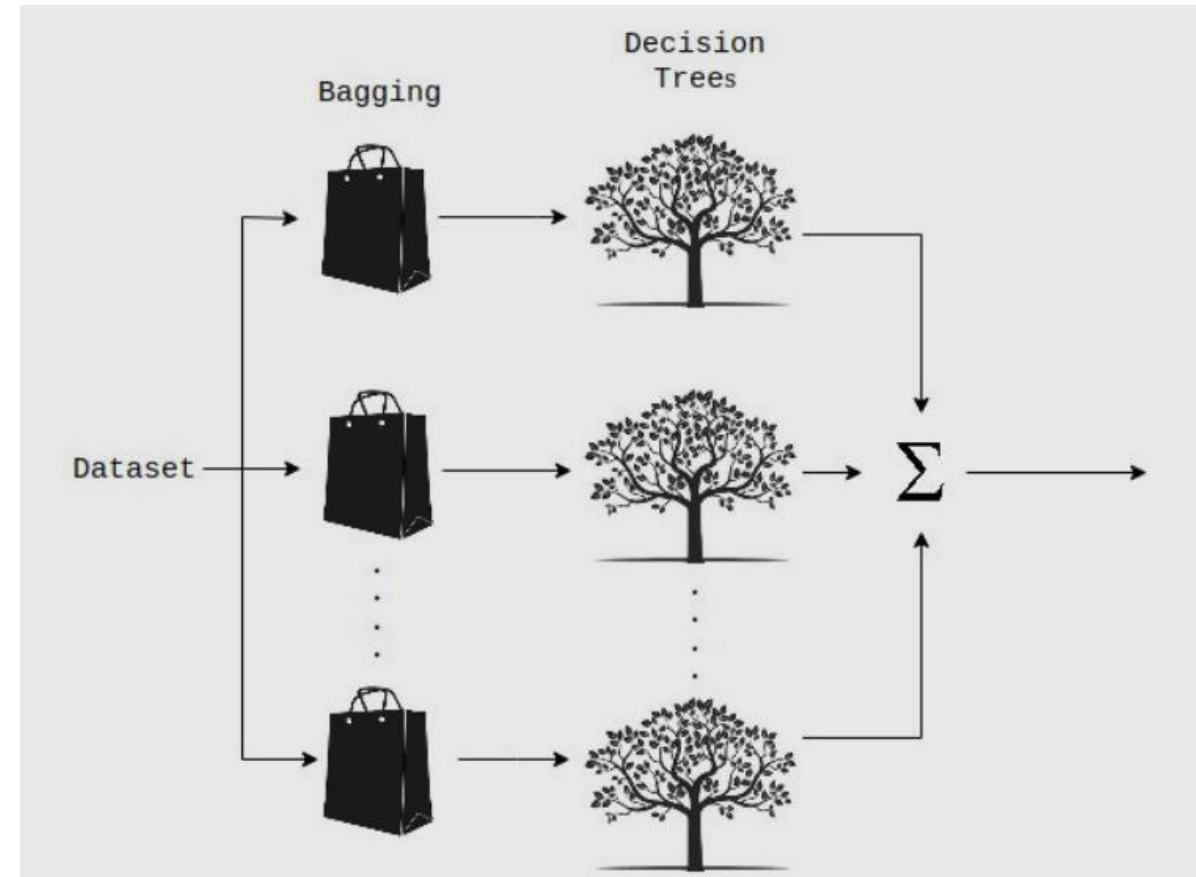
THE DATASET

- The dataset we used for this project is a collection of prices of different vehicles available in Indian markets.
- The dataset is made of total 302 rows and 10 columns.
- The type of variables in this dataset are described in the table.

Variable	Description
Car Name	Names of the vehicles
Year	Year of purchase
Selling Price	The prices these vehicles were sold for
Present Price	The prices these vehicles were bought for
Kms Driven	Total kilometers driven by the vehicle
Fuel Type	The type of fuel a vehicle used (Petrol, Diesel or CNG)
Seller Type	Who is selling the vehicle (Dealer or Individual)
Transmission	Transmission system of the vehicle (Manual or Automatic)
Owner	Count of previous owners of a particular vehicle

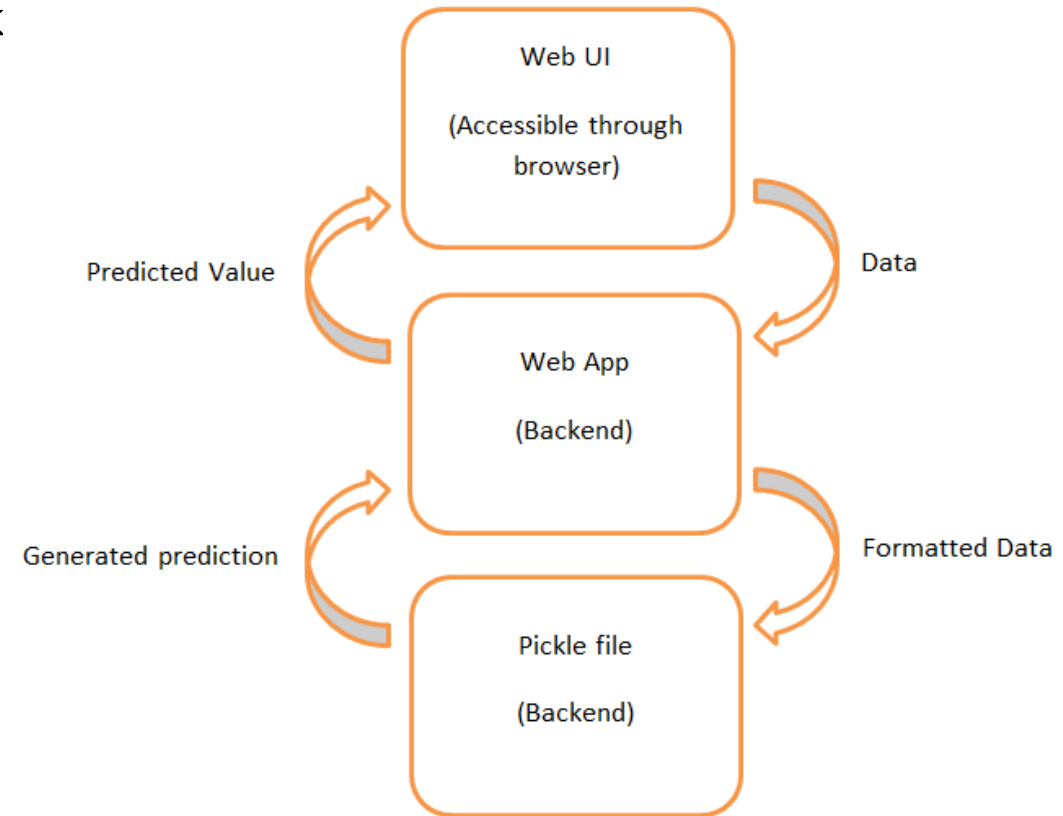
THE ML ALGORITHM: RANDOM FOREST REGRESSOR

- It is an ensemble technique capable of performing both regression and classification
- Random Forests build multiple decision trees and merge their predictions together to get a more accurate prediction
- The technique of combining predictions of n different models is called bagging



THE WEBAPP

- The Web Application is developed using Python Flask Framework
- The user is required to enter data through the Web UI
- The raw data gets sent to the WebApp
- It then is converted into required format
- The WebApp then opens the Pickle file and feeds the data
- The .pkl file then computes the data and generated a prediction and returns the predicted value to the WebApp
- The WebApp then displays this predicted value on the UI

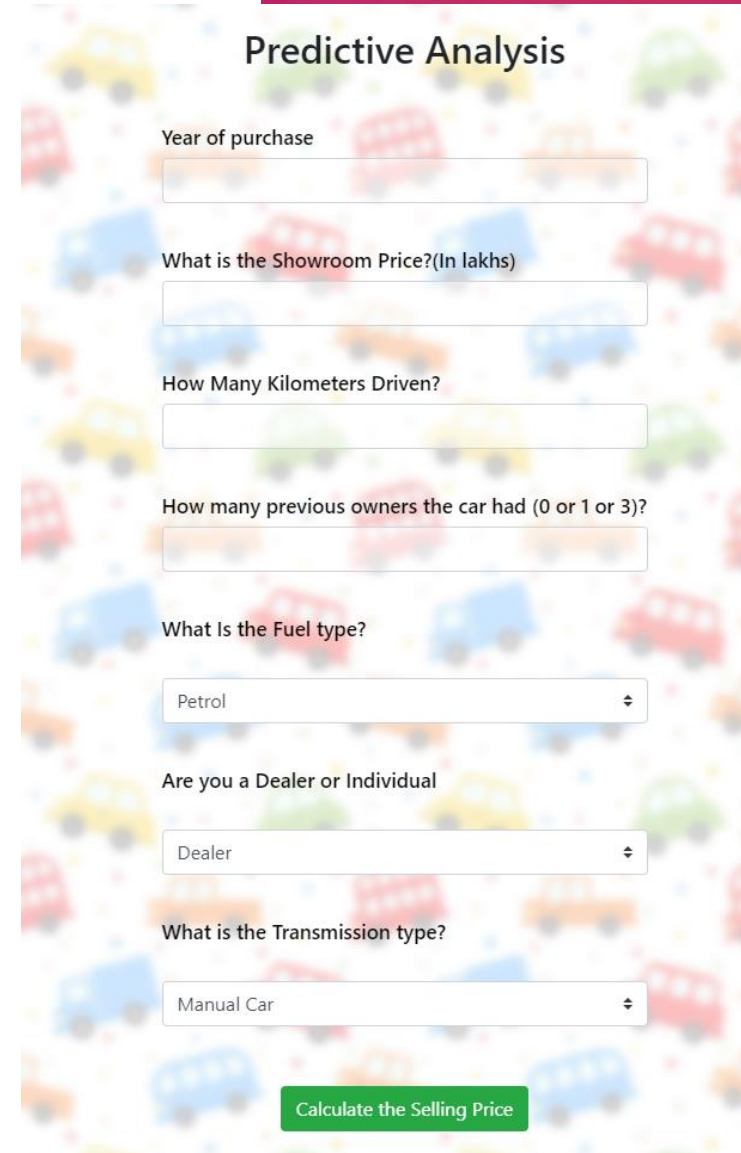


DEPLOYING THE WEBAPP

- Deploying the WebApp was the last phase of the project
- The project is deployed on Heroku using GitHub repository for hosting files.

Why we used Heroku?

- It is free to use
- Deploying is easy on Heroku
- Environments are easy to maintain
- It supports Python as backend

A web form titled "Predictive Analysis" for calculating a car's selling price. The form is set against a background of colorful cartoon cars. It contains several input fields and dropdown menus. At the bottom is a green button labeled "Calculate the Selling Price".

Predictive Analysis

Year of purchase

What is the Showroom Price?(In lakhs)

How Many Kilometers Driven?

How many previous owners the car had (0 or 1 or 3)?

What Is the Fuel type?

Are you a Dealer or Individual

What is the Transmission type?

CONCLUSION

- The aim of this project was to successfully demonstrate a solution for the most common problem faced in developing an ML Model which is deployment of an application that is simple enough for users to access and use.
- We conclude that the Random Forest Regression model was a right fit for the problem statement and for the type of dataset that we have used. And we have successfully developed an Web Application for the project. The same project can be scaled for new features and larger datasets as it is modular and scalable.

REFERENCES

- [Scikit Learn Random Forest Regressor](#)
- [Seaborn Data Visualization Library](#)
- [Matplotlib Data Visualization Library](#)
- [Dataset used for the project](#)
- [Random Forest Regressor](#)

END