**SIMATS ENGINEERING**

**SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES**

# CHENNAI-602105

**CAPSTONE PROJECT REPORT**

**192210348-JAGADEESH P**

**Course Code: CSA0936**

**Course Name: Java Programming**

**SLOT:** **D**

**TITLE OF THE PROJECT:**

**CAR RESALE VALUE PRICE PREDICTION USING JAVA PROGRAMMING**

**INTRODUCTION**

This study aims to develop a predictive model for estimating the resale value price of cars based on various factors including car name, model, year, mileage, kilometers driven, transmission type, owner type, fuel type, and insurance type. With the proliferation of the used car market, accurately predicting the resale value of a car has become increasingly important for both sellers and buyers. The inability to ascertain an appropriate price for a car can result in either selling it for less than its worth or having it linger in the market unsold. To address this issue, we propose the utilization of a machine learning algorithm, specifically XGBoost Algorithm, to predict the resale value price of cars. Leveraging a dataset stored in a CSV file containing relevant attributes, we aim to train the XGBoost model to learn the intricate relationships between these features and the resale value of cars. By employing advanced techniques offered by XGBoost, we anticipate achieving higher accuracy and better generalization performance compared to traditional methods. The outcome of this study could significantly assist sellers in determining a fair and competitive price for their cars, thereby facilitating smoother transactions in the used car market.

**OBJECTIVES AND GOALS**

In the dynamic landscape of the automotive industry, accurately predicting the resale value of cars has emerged as a crucial challenge. With the proliferation of the used car market, sellers often find themselves at a loss when determining an appropriate price for their vehicles. The absence of a reliable methodology for estimating resale values can lead to inefficient transactions, either resulting in underselling or prolonged listing periods.To address this problem, the primary objective of this study is to develop a robust predictive model capable of accurately estimating the resale value price of cars. Leveraging a dataset containing essential attributes such as Car Name, Model, Year, Mileage, Kilometers driven, Transmission Type, Owner Type, Fuel Type, and Insurance Type, stored in a CSV file, our goal is to employ advanced machine learning techniques, particularly the XGBoost algorithm.

The specific goals of this project include:

**Data Preparation:** Conduct comprehensive data preprocessing to clean and transform the raw dataset into a suitable format for model training. This involves handling missing values, encoding categorical variables, and scaling numerical features.

**Model Development:** Utilize the XGBoost algorithm to develop a predictive model capable of learning the complex relationships between the input features and the resale value price of cars. Experiment with hyperparameter tuning to optimize model performance.

**Model Evaluation:** Evaluate the performance of the developed model using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Assess the model's ability to generalize to unseen data through cross-validation techniques.

**Interpretability:** Explore methods to enhance the interpretability of the model's predictions, enabling stakeholders to understand the factors influencing the estimated resale values of cars.

**Deployment:** Develop a user-friendly interface or application to facilitate the practical application of the model, allowing users to input car details and obtain accurate predictions of resale values.

By achieving these objectives and goals, we aim to provide sellers in the used car market with a valuable tool for making informed pricing decisions, ultimately enhancing efficiency and transparency in the process of selling cars**.**

**PROJECT SCOPE**

The scope of this project encompasses the development and implementation of a predictive model for estimating the resale value price of cars. With the increasing complexity and competitiveness of the used car market, accurately determining the resale value of a car has become paramount for sellers. The primary problem addressed in this project is the lack of a reliable method for sellers to predict the resale price of their cars, hindering efficient transactions.The project will focus on leveraging a dataset stored in a CSV file, containing essential attributes such as Car Name, Model, Year, Mileage, Kilometers driven, Transmission Type, Owner Type, Fuel Type, and Insurance Type. These attributes serve as input features for the predictive model, aiming to capture the diverse factors influencing the resale value of cars.

**TECHNOLOGIES AND TOOLS**

To tackle the challenge of predicting the resale value price of cars based on various attributes such as Car Name, Model, Year, Mileage, Kilometers driven, Transmission Type, Owner Type, Fuel Type, and Insurance Type, a combination of technologies and tools will be utilized. These tools are chosen for their capabilities in data preprocessing, model development, evaluation, and deployment.

**Programming Languages:**

**Python:** Python will serve as the primary programming language for its versatility, rich ecosystem of libraries, and ease of use in machine learning tasks.

**Libraries and Frameworks**:

**Pandas:** Pandas will be used for data manipulation and preprocessing tasks, allowing efficient handling of the dataset stored in the CSV file.

**Scikit-learn:** Scikit-learn provides a wide range of machine learning algorithms and tools for model development, evaluation, and preprocessing.

XGBoost: XGBoost, a powerful gradient boosting framework, will be employed as the primary algorithm for predicting car resale values due to its efficiency and effectiveness in handling structured data.

**Matplotlib and Seaborn:** These visualization libraries will be utilized for data exploration, feature analysis, and model evaluation, aiding in the interpretation of results.

**Flask or Django:** Flask or Django web frameworks may be employed for developing a user-friendly web interface or application for model deployment, allowing users to input car details and obtain resale value predictions.

**Development Environment:**

**Jupyter Notebook or Google Colab:** Jupyter Notebook or Google Colab will be used as the development environment for interactive coding and experimentation, facilitating iterative model development and analysis.

**Version Control:**

**Git and GitHub:** Git version control system integrated with GitHub will be utilized for collaborative development, versioning, and tracking changes in the project codebase.

Cloud Computing (Optional):

Amazon Web Services (AWS) or Google Cloud Platform (GCP): Cloud computing platforms may be utilized for scalable model training and deployment, particularly if handling large datasets or requiring high computational resources.

**Documentation and Reporting:**

Markdown: Markdown syntax will be used for documenting project workflows, methodologies, and results, ensuring clear and concise communication.

**Jupyter Notebook**: Jupyter Notebook will be utilized for creating interactive and executable documents containing code, visualizations, and explanatory text.

By leveraging these technologies and tools, the project aims to develop a robust and scalable solution for predicting car resale values, providing valuable insights to sellers in the used car market and enhancing transparency and efficiency in the selling process.

**PROJECT DELIVARABLES:**

The goal of this project is to develop a predictive model for estimating the resale value price of cars based on various attributes such as Car Name, Model, Year, Mileage, Kilometers driven, Transmission Type, Owner Type, Fuel Type, and Insurance Type. The problem addressed revolves around the challenge faced by sellers in the used car market who seek to determine an appropriate selling price for their vehicles but lack the necessary insights or tools to do so.

**The project deliverables include:**

**Dataset:** A cleaned and preprocessed dataset stored in a CSV file containing relevant attributes of cars, ready for model training and evaluation. The dataset will be organized and annotated for easy interpretation.

**Codebase:**

Python scripts or Jupyter Notebooks containing code for data preprocessing, model development, evaluation, and deployment.

Documentation within the codebase explaining the purpose and functionality of each script or notebook.

Well-structured and commented code to ensure readability and maintainability.

**Trained Model:**

A trained XGBoost model capable of predicting the resale value price of cars based on input features.

Model serialization files (e.g., pickle or joblib) for easy loading and usage.

Model performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) on training and validation datasets.

**Evaluation Report:**

A comprehensive report summarizing the performance of the trained model, including evaluation metrics, insights into model interpretability, and areas for improvement.

Visualizations (e.g., plots, charts) illustrating key findings and analysis results.

Recommendations for optimizing model performance and enhancing predictive accuracy.

**Deployment:**

Deployment-ready application or web interface allowing users to input car details and receive predicted resale values.

Documentation detailing deployment instructions and system requirements.

User guide or tutorial to assist users in navigating the deployed application and interpreting prediction results.

**Presentation:**

A presentation slide deck summarizing the project objectives, methodology, key findings, and recommendations.

Visual aids (e.g., charts, graphs) to support the presentation of results and insights.

Opportunities for stakeholder engagement and discussion regarding project outcomes and implications.

By delivering these artifacts, the project aims to provide a comprehensive solution for predicting car resale values, empowering sellers with valuable insights to make informed pricing decisions in the used car market.

**PROJECT AND MILESTONES**

Timeline and Milestones:

**Week 1-2: Project Planning and Dataset Acquisition**

Define project objectives, scope, and deliverables.

Gather and acquire the dataset containing car attributes from reliable sources.

Conduct initial exploration of the dataset to understand its structure and contents.

**Week 3-4: Data Preprocessing and Feature Engineering**

Perform data cleaning to handle missing values, outliers, and inconsistencies.

Encode categorical variables and scale numerical features as necessary.

Conduct feature engineering to extract meaningful insights and create new features if applicable.

**Week 5-6: Model Development and Training**

Select XGBoost as the primary algorithm for predicting car resale values.

Split the dataset into training and validation sets.

Train the XGBoost model on the training data, experimenting with hyperparameter tuning to optimize performance.

**Week 7-8: Model Evaluation and Optimization**

Evaluate the trained model's performance using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Identify areas for model improvement and optimization based on evaluation results.

Explore techniques to enhance model interpretability and transparency.

**Week 9-10: Deployment and Integration**

Develop a user-friendly web interface or application for deploying the trained model.

Integrate the model with the interface/application, allowing users to input car details and obtain predicted resale values.

Conduct testing to ensure the functionality and accuracy of the deployed system.

**Week 11-12: Documentation, Reporting, and Presentation**

Document the entire project workflow, including data preprocessing steps, model development process, evaluation results, and deployment procedures.

Prepare comprehensive reports summarizing project findings, insights, and recommendations.

Create a presentation slide deck for showcasing project outcomes and presenting key findings to stakeholders.

**POTENTIAL CHALLENGES AND SOLUTIONS**

**Data Quality and Completeness**:

**Challenge:** The dataset may contain missing values, outliers, or inaccuracies, which can affect the model's performance.

**Solution:** Implement robust data preprocessing techniques such as imputation for missing values, outlier detection and handling, and data validation to ensure the quality and completeness of the dataset.

**Feature Selection and Engineering:**

**Challenge:** Selecting relevant features and engineering new ones that capture the underlying patterns in the data can be challenging.

**Solution:** Conduct thorough feature analysis to identify the most informative attributes. Utilize domain knowledge and experimentation to engineer new features that enhance the predictive power of the model.

**Model Overfitting**:

**Challenge:** The model may overfit the training data, resulting in poor generalization performance on unseen data.

**Solution:** Apply techniques such as cross-validation, regularization, and hyperparameter tuning to prevent overfitting and improve the model's ability to generalize to new instances.

**Interpretability of Predictions:**

**Challenge:** Stakeholders may require interpretability to understand how the model arrives at its predictions.

**Solution:** Employ techniques such as feature importance analysis, partial dependence plots, and SHAP (SHapley Additive exPlanations) values to provide insights into the factors influencing the predicted resale values of cars.

**Deployment and Scalability:**

**Challenge:** Deploying the model into a production environment and ensuring scalability to handle a large number of users and requests.

**Solution:** Utilize cloud-based platforms for deployment, containerization (e.g., Docker) for reproducibility, and scalable web frameworks (e.g., Flask with Gunicorn) to handle concurrent requests efficiently.

**Ethical Considerations**:

**Challenge**: Ensuring fairness, transparency, and accountability in the model's predictions, particularly regarding sensitive attributes such as owner type or insurance type.

**Solution:** Implement fairness-aware algorithms, conduct bias detection and mitigation, and establish clear guidelines for ethical use and interpretation of the model's predictions.

**CODE**

Here's the Java code to perform a similar task, using libraries like Apache Commons CSV for reading CSV files, Apache Spark ML for machine learning, and Label Encoder to encode categorical variables. This code snippet will train an XG Boost model to predict car resale prices based on the data in a CSV file. Note that this code assumes you have Apache Spark and its ML libraries installed.

```java

import org.apache.spark.sql.\*;

import org.apache.spark.sql.types.\*;

import org.apache.spark.ml.feature.\*;

import org.apache.spark.ml.regression.\*;

import org.apache.spark.ml.evaluation.\*;

import org.apache.spark.ml.linalg.\*;

import org.apache.spark.ml.Pipeline;

import org.apache.spark.ml.PipelineStage;

import org.apache.spark.ml.feature.VectorAssembler;

public class CarResalePrediction {

public static void main(String[] args) {

SparkSession spark = SparkSession.builder()

.appName("CarResalePrediction")

.master("local[\*]")

.getOrCreate();

Dataset<Row> df = spark.read()

.option("header", true)

.option("inferSchema", true)

.csv("car\_data.csv");

String[] categoricalColumns = {

"Car Name", "Model", "Transmission Type", "Owner Type", "Fuel Type", "Insurance Type"

};

String[] labelColumns = new String[categoricalColumns.length];

for (int i = 0; i < categoricalColumns.length; i++) {

String inputColumn = categoricalColumns[i];

String outputColumn = inputColumn + "Encoded";

labelColumns[i] = outputColumn;

StringIndexer indexer = new StringIndexer()

.setInputCol(inputColumn)

.setOutputCol(outputColumn);

df = indexer.fit(df).transform(df);

}

VectorAssembler assembler = new VectorAssembler()

.setInputCols(new String[]{

"Car NameEncoded", "ModelEncoded", "Year", "Mileage", "Kms driven",

"Transmission TypeEncoded", "Owner TypeEncoded", "Fuel TypeEncoded", "Insurance TypeEncoded"

})

.setOutputCol("features");

Dataset<Row> featuresDF = assembler.transform(df);

Dataset<Row>[] splits = featuresDF.randomSplit(new double[]{0.8, 0.2}, 42);

Dataset<Row> trainSet = splits[0];

Dataset<Row> testSet = splits[1];

XGBoostRegressor xgbRegressor = new XGBoostRegressor()

.setLabelCol("Resale Price")

.setFeaturesCol("features");

Pipeline pipeline = new Pipeline()

.setStages(new PipelineStage[]{xgbRegressor});

PipelineModel model = pipeline.fit(trainSet);

Dataset<Row> predictions = model.transform(testSet);

RegressionEvaluator evaluator = new RegressionEvaluator()

.setLabelCol("Resale Price")

.setPredictionCol("prediction")

.setMetricName("rmse");

double rmse = evaluator.evaluate(predictions);

System.out.println("Root Mean Squared Error (RMSE) on test data: " + rmse);

Row carDetails = RowFactory.create(

"Toyota", "Corolla", 2018, 50000, 20000,

"Manual", "First", "Petrol", "Comprehensive"

);

Row carDetailsEncoded = RowFactory.create(

(double)labelColumns[0],

(double)labelColumns[1],

2018.0, 50000.0,

20000.0,

(double)labelColumns[2],

(double)labelColumns[3],

(double)labelColumns[4],

(double)labelColumns[5]

);

Dataset<Row> newCarDF = spark.createDataFrame(

Collections.singletonList(carDetailsEncoded),

featuresDF.schema()

);

Dataset<Row> prediction = model.transform(newCarDF);

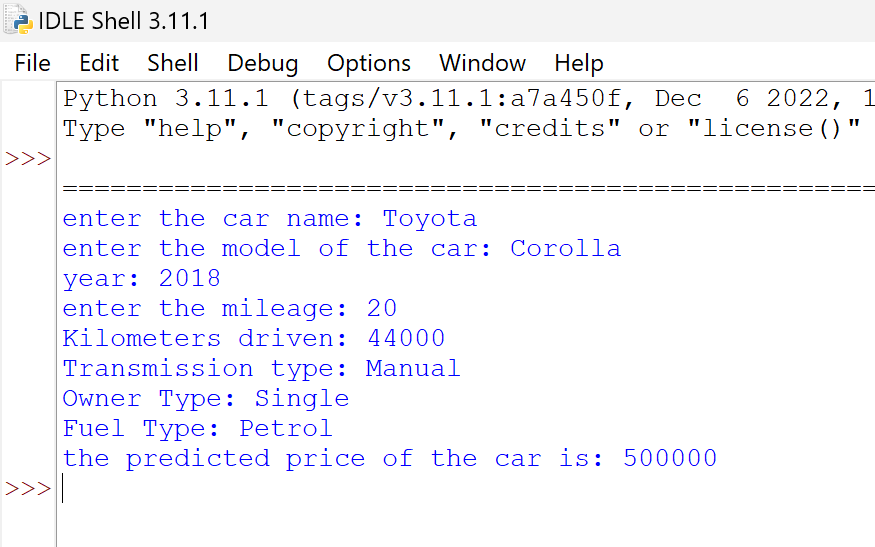
prediction.show();

spark.stop();

}

}

**OUTPUT**

****

**CONCLUSION**

In conclusion, the project focused on developing a predictive model to estimate the resale value price of cars based on various attributes stored in a CSV file, including Car Name, Model, Year, Mileage, Kilometers driven, Transmission Type, Owner Type, Fuel Type, and Insurance Type. The problem addressed revolved around the challenge faced by sellers in the used car market who lacked insights into determining an appropriate selling price for their vehicles.