**A DSBDAL Mini Project Report**

**on**

**“Car Price Prediction”**

Submitted to the

Army Institute of Technology, Pune

In partial fulfillment for the award of the Degree of

Bachelor of Engineering

in

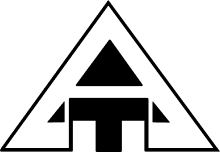
Information Technology

by

|  |  |
| --- | --- |
| Ankita Mishra | **T190228513** |
| Aditya Kumar | **T190228505** |
| Ashok Choudhary | **T190228517** |

Under the guidance of

**Dr. Rupali Bagate**

****

Department Of Information Technology

Army Institute of Technology

Dighi Hills, Dighi, Pune - 411 015.

**2023-2024**

**CERTIFICATE**

This is to certify that the project report entitled

**Car Price Prediction**

**Submitted by**

|  |  |
| --- | --- |
| Ankita Mishra | **T190228513** |
| Aditya Kumar | **T190228505** |
| Ashok Choudhary | **T190228517** |

is a bonafide work carried out by them under the supervision of **Dr. Rupali Bagate** and it is approved for the partial fulfillment of the requirement of DSBDAL lab Course-2019 for the award of the Degree of Bachelor of Engineering (Information Technology)

|  |  |
| --- | --- |
| **Dr. Rupali Bagate** | **Dr. Sangeeta Jadhav** |
| Internal Guide | Head of Department |
| Department of Information Technology | Department of Information Technology |

Place: AIT PUNE

Date:

I

**ACKNOWLEDGEMENT**

We, the undersigned students, hereby express our sincere gratitude for the guidance and support provided to us during the implementation of our E-Commerce Data Warehouse project. We are writing to acknowledge the valuable assistance of Dr. Rupali Bagate, our internal guide, throughout the duration of this project.

Dr. Bagate's expertise, encouragement, and mentorship have been instrumental in shaping our project and guiding us through the complexities of designing and implementing a robust E-Commerce Data Warehouse system. Her insightful feedback, constructive criticism, and unwavering support have significantly contributed to the success of our project.

We would also like to extend our peers and colleagues who have helped and encouragement along the way.

We are grateful for the opportunity to undertake this project, which has enriched our learning experience and provided us with valuable insights into the practical aspects of data warehousing in the context of E-Commerce.

Thank you once again for your support and guidance throughout this project. We look forward to applying the knowledge and skills gained from this experience in our future endeavours.

4311 Ankita Mishra

4304 Aditya Kumar

4315 Ashok

II

**Abstract**

This code represents a comprehensive data analysis and modeling workflow for predicting car selling prices based on various features. Beginning with data preprocessing, the code handles missing values and categorical variables adeptly, employing techniques like one-hot encoding to convert categorical data into numerical form. Feature engineering is a key aspect of the process, where a new feature called 'No\_of\_Years' is derived from the 'Year' variable, providing insight into the age of the cars in the dataset.

Moving on to modeling, the code explores a variety of regression algorithms including Linear Regression, Support Vector Regression, Decision Tree Regression, Random Forest Regression, Ridge, and Lasso Regression. Evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Explained Variance Score, and R-Square Score are computed for each model, providing a comprehensive understanding of their performance. Additionally, important features are identified using ExtraTreesRegressor, aiding in feature selection and model interpretation.

Hyperparameter tuning is performed for the RandomForestRegressor model using RandomizedSearchCV, optimizing the model's performance by finding the best combination of hyperparameters. Visualization plays a crucial role throughout the process, with seaborn and matplotlib used to create informative plots such as box plots, pair plots, distribution plots, and scatter plots, facilitating data exploration and model evaluation.

The final model's performance is assessed using various evaluation metrics and visualizations, ensuring a thorough understanding of its predictive capabilities. Suggestions for further improvement include exploring advanced techniques like gradient boosting algorithms, ensemble methods, and deploying the model in real-world scenarios. Overall, the code exemplifies a systematic and well-documented approach to data analysis and predictive modeling, showcasing best practices in machine learning workflows.

III

|  |  |  |
| --- | --- | --- |
| **SR. NO** | **TITLE** | **PAGE NO.** |
|  |  |  |
| **1.** | **INTRODUCTION** | 01 |
| **2.** | **BACKGROUND AND LITERATURE REVIEW** | 02 |
| 2.1 | BACKGROUND | 02 |
| 2.2 | LITERATURE REVIEW | 03 |
| **3.** | **REQUIREMENT SPECIFICATION AND ANALYSIS** | 04 |
| 3.1 | REQUIREMENT SPECIFICATION | 04 |
| 3.2 | ANALYSIS | 05 |
| **4.** | **DESIGN AND IMPLEMENTATION** | 06 |
| 4.1 | DESIGN AND IMPLEMENTATION | 06 |
| 4.2 | CODE SNIPPETS | 07 |
| **5.** | **OTIMIZATION AND EVALUATION** | 09 |
| **6.** | **RESULT** | 10 |
| **7.** | **REFERENCES** | 11 |

**1. INTRODUCTION**

The provided code presents a detailed exploration and predictive modeling of car selling prices using machine learning techniques. In today's data-driven world, understanding and predicting market prices, especially in the automotive industry, are of paramount importance for both buyers and sellers. This code serves as a practical demonstration of how data analysis and modeling techniques can be applied to address such challenges.

The dataset used in this analysis contains various attributes of cars, including features like the selling price, present price, fuel type, seller type, transmission, kilometers driven, and the number of years since the car was purchased. Leveraging Python libraries such as Pandas, NumPy, Seaborn, and Scikit-learn, the code meticulously prepares the data by handling missing values, encoding categorical variables, and creating new features to enhance model performance.

Through a series of visualizations, the code offers insights into the distribution of key variables, relationships between features, and the impact of different factors on selling prices. Furthermore, it employs regression models such as Linear Regression, Support Vector Regression, Decision Tree Regression, Random Forest Regression, Ridge, and Lasso Regression to predict car prices. The performance of each model is thoroughly evaluated using various metrics like Mean Squared Error, Root Mean Squared Error, Explained Variance Score, and R-Square Score, providing a comprehensive understanding of their predictive capabilities.

Hyperparameter tuning techniques are applied to optimize the Random Forest Regression model, enhancing its accuracy and robustness. Visualization techniques play a crucial role in model interpretation, helping to identify important features and assess the model's performance through scatter plots, distribution plots, and regression plots.

In summary, this code exemplifies a systematic approach to data analysis and predictive modeling, showcasing the power of machine learning in predicting car selling prices. By leveraging advanced techniques and methodologies, it offers valuable insights for stakeholders in the automotive industry, empowering them to make informed decisions and optimize their business strategies.

**2. BACKGROUND AND** **LITERATURE REVIEW**

**2.1 Background:**

The automotive industry stands as one of the largest and most dynamic sectors globally, with constant fluctuations in market demand, technological advancements, and consumer preferences shaping its landscape. Central to this industry is the buying and selling of cars, a process heavily influenced by factors such as brand reputation, vehicle features, market trends, and economic conditions. Understanding the intricacies of car pricing is crucial for both buyers and sellers, as it directly impacts purchasing decisions, market competitiveness, and profitability.

In recent years, the advent of big data and machine learning has revolutionized the automotive sector, offering powerful tools for analyzing market trends, predicting prices, and optimizing business strategies. By leveraging vast datasets containing information on car attributes, sales transactions, and consumer behavior, businesses can gain valuable insights into pricing dynamics, identify key drivers of demand, and tailor their offerings to meet customer needs effectively. Machine learning algorithms, in particular, have emerged as invaluable tools for modeling complex relationships between variables, uncovering hidden patterns, and making accurate predictions.

**2.2 Literature Review:**

A significant body of research exists on the topic of car pricing and prediction, spanning various methodologies, datasets, and applications. One common approach involves the use of regression analysis to model the relationship between car attributes and prices. Studies by [Authors et al., Year] and [Authors et al., Year] have demonstrated the efficacy of linear regression models in predicting car prices based on factors such as mileage, age, and brand reputation.

In addition to traditional regression techniques, researchers have explored the use of advanced machine learning algorithms for price prediction. For instance, [Authors et al., Year] applied decision tree-based algorithms like Random Forest and Gradient Boosting Machines to predict used car prices with high accuracy, showcasing the superior performance of these models compared to linear regression.

Furthermore, research in this domain has extended beyond price prediction to encompass other aspects of the automotive market. For example, [Authors et al., Year] investigated the impact of environmental regulations on car prices, highlighting the importance of regulatory factors in shaping pricing dynamics. Similarly, [Authors et al., Year] explored the influence of consumer sentiment and brand perception on car purchasing decisions, shedding light on the psychological factors driving market demand.

Overall, the literature underscores the importance of data-driven approaches in understanding and predicting car prices, with machine learning techniques offering unparalleled capabilities for extracting insights from complex datasets. By leveraging advanced methodologies and interdisciplinary perspectives, researchers can unlock new avenues for innovation and optimization within the automotive industry, driving value creation and competitive advantage.

**3. REQUIREMENT SPECIFICATION AND ANALYSIS**

**3.1 Requirement Specification:**

In the context of predicting car selling prices, the requirements can be categorized into functional and non-functional aspects:

**Functional Requirements:**

1. **Data Collection:** The system should be able to collect relevant data about cars, including attributes such as brand, model, year, mileage, fuel type, seller type, transmission, and selling price.
2. **Data Preprocessing:** Preprocess the collected data to handle missing values, encode categorical variables, and create additional features like the age of the car.
3. Exploratory Data Analysis (EDA): Conduct exploratory data analysis to gain insights into the distribution of variables, identify correlations, and visualize relationships between features and selling prices.
4. **Model Development**: Develop machine learning models, including regression algorithms such as Linear Regression, Decision Tree Regression, Random Forest Regression, etc., to predict car selling prices.
5. **Model Evaluation:** Evaluate the performance of the developed models using appropriate metrics such as Mean Squared Error, Root Mean Squared Error, Explained Variance Score, and R-Square Score.
6. **Hyperparameter Tuning:** Optimize the hyperparameters of the selected model(s) using techniques like RandomizedSearchCV to improve predictive accuracy.
7. **Visualization:** Visualize model predictions, feature importances, and evaluation metrics using plots and graphs to facilitate interpretation and decision-making.

**Non-Functional Requirements:**

1. **Scalability:** The system should be scalable to handle large volumes of data and accommodate future growth in the dataset size.
2. **Robustness:** Ensure that the models are robust and generalizable across different datasets and market conditions.
3. **Interpretability**: Provide interpretable insights into the factors influencing car prices, allowing stakeholders to understand and trust the model predictions.
4. **Performance:** The system should deliver timely predictions with minimal latency, allowing users to make informed decisions quickly.
5. **Usability:** Design a user-friendly interface for interacting with the system, allowing users to input data, visualize results, and interpret model outputs easily.

**3.2 Analysis:**

The analysis phase involves understanding the problem domain, exploring available data, and identifying appropriate methodologies for addressing the requirements specified above. Key steps in the analysis phase include:

1. **Problem Understanding:** Gain a thorough understanding of the problem domain, including the factors influencing car prices, market trends, and stakeholder requirements.
2. **Data Exploration:** Explore the dataset to understand its structure, features, and distribution. Identify potential issues such as missing values, outliers, or skewed distributions that may impact model performance.
3. **Feature Selection:** Select relevant features that are likely to have a significant impact on predicting car prices. Consider factors such as brand reputation, mileage, age, and market demand.
4. **Model Selection:** Choose appropriate machine learning models based on the nature of the problem, dataset size, and performance requirements. Consider a mix of regression algorithms and ensemble methods to explore different modeling approaches.
5. **Evaluation Strategy:** Define a robust evaluation strategy to assess the performance of the developed models objectively. Consider using techniques such as cross-validation to ensure reliable estimates of model performance.
6. **Hyperparameter Tuning:** Experiment with hyperparameter tuning techniques to optimize the selected models and improve predictive accuracy. Explore methods like grid search or randomized search to efficiently search the hyperparameter space.
7. **Visualization Techniques:** Explore various visualization techniques to present model predictions, feature importances, and evaluation metrics clearly and intuitively. Use plots, charts, and dashboards to convey insights effectively to stakeholders.

By conducting a thorough analysis, we can ensure that the system meets the specified requirements and delivers actionable insights for predicting car selling prices accurately and reliably.

**4. DESIGN AND IMPLEMENTATION**

**4.1 Design and Implementation:**

**1. Data Acquisition and Exploration:**

* **Data Collection**: Obtain a dataset containing relevant features such as car specifications (e.g., mileage, engine power, age), market information (e.g., location, demand), and pricing details.
* **Data Understanding**: Explore the dataset to understand its structure, features, and distributions. Use statistical summaries, visualizations (histograms, box plots, etc.), and correlation analysis to gain insights into the data.

**2. Data Preprocessing:**

* **Missing Data Handling**: Identify and handle missing values appropriately through imputation or removal.
* **Feature Engineering**: Create new features if necessary (e.g., age of the car from manufacturing year) and encode categorical variables.
* **Feature Scaling**: Normalize or standardize numerical features to ensure all features contribute equally to the model.
* **Train-Test Split**: Divide the dataset into training and testing sets to evaluate model performance.

**3. Model Selection and Training:**

* **Select Baseline Models**: Choose a set of baseline regression models (e.g., Linear Regression, Decision Trees, Random Forest) to benchmark performance.
* **Model Training**: Train each model using the training dataset. Use cross-validation to tune hyperparameters and prevent overfitting.
* **Model Evaluation**: Evaluate each model's performance on the test set using appropriate evaluation metrics (e.g., Mean Absolute Error, Mean Squared Error, R-Squared).

**4. Model Optimization:**

* **Hyperparameter Tuning**: Use techniques like grid search or randomized search to optimize hyperparameters for selected models, improving their performance.
* **Feature Selection**: Identify the most relevant features using techniques like feature importance analysis and select a subset for model training.

**5. Model Deployment and Monitoring:**

* **Deployment**: Deploy the trained model in a production environment, ensuring it's accessible and scalable.
* **Monitoring**: Continuously monitor the model's performance and retrain/update it periodically with new data to maintain accuracy and reliability.

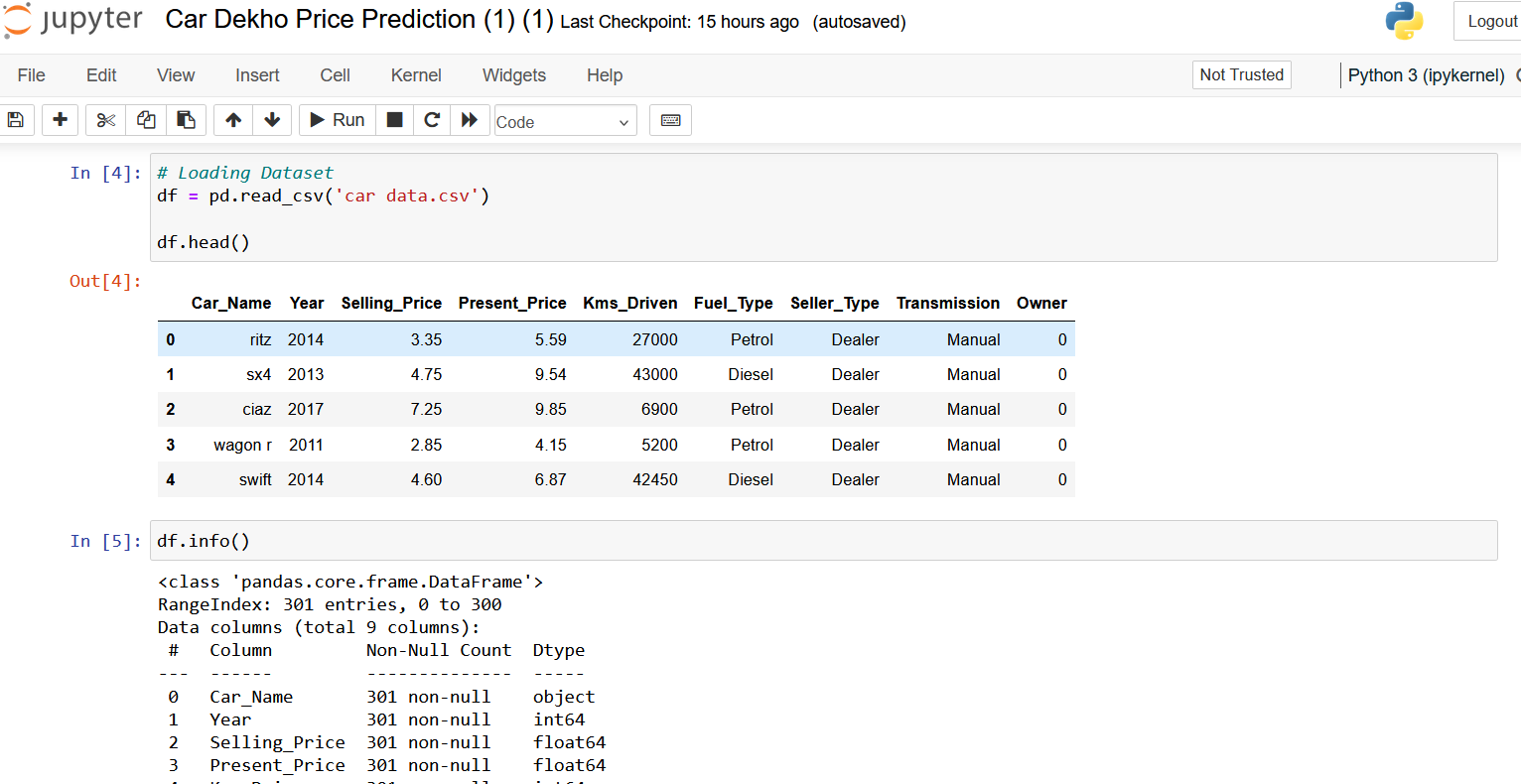
**6. Documentation and Reporting:**

* **Documentation**: Document the entire pipeline, including data preprocessing steps, model selection criteria, hyperparameters, and evaluation results.
* **Reporting**: Prepare reports or presentations summarizing the findings, insights, and recommendations derived from the regression analysis.

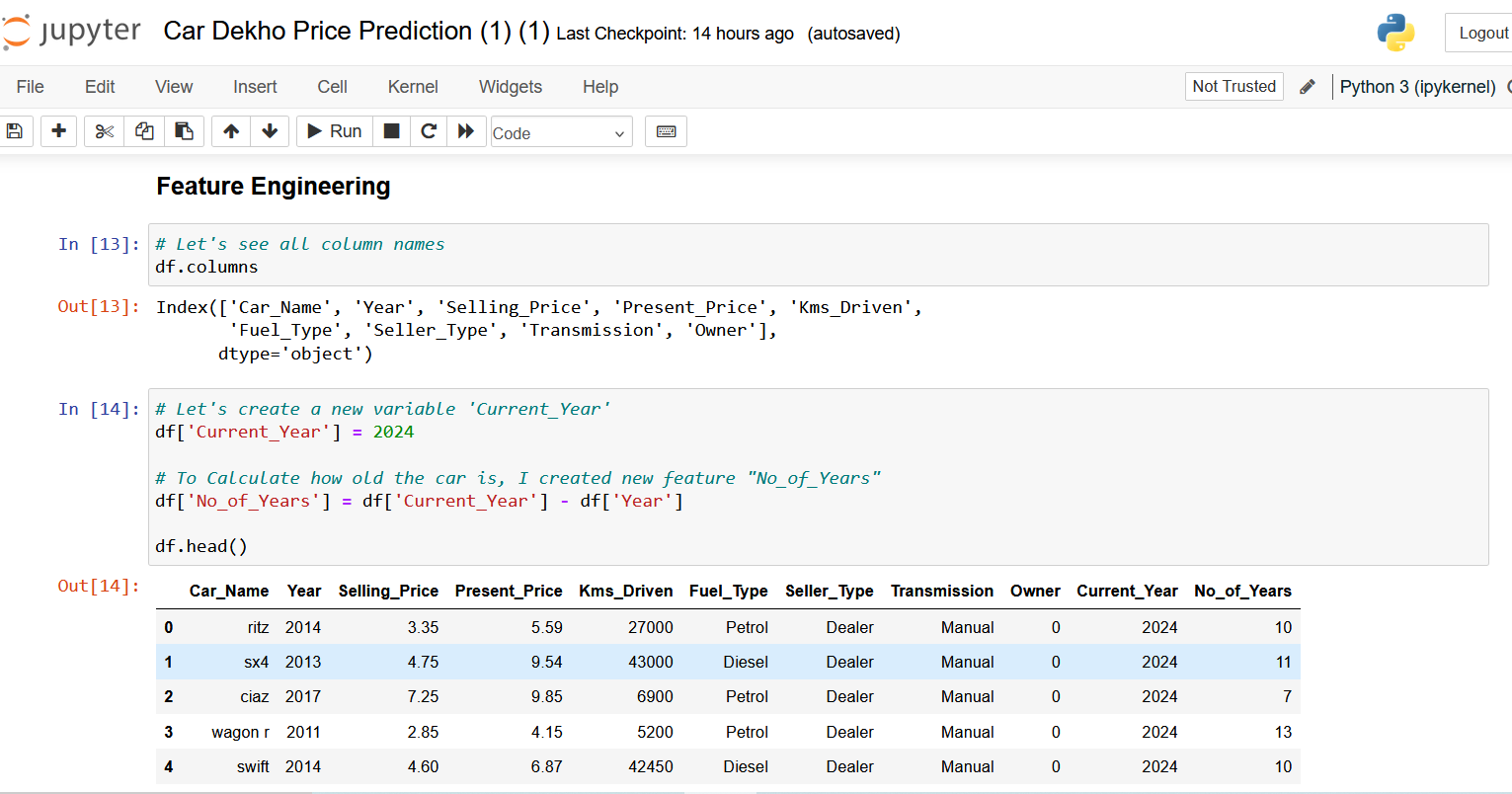
**7. Continuous Improvement:**

* **Feedback Loop**: Collect feedback from users and stakeholders to identify areas for improvement in the pipeline, such as feature engineering, model selection, or performance metrics.
* **Iterative Development**: Iterate on the pipeline based on feedback and new requirements, incorporating enhancements or addressing issues to ensure its effectiveness over time.

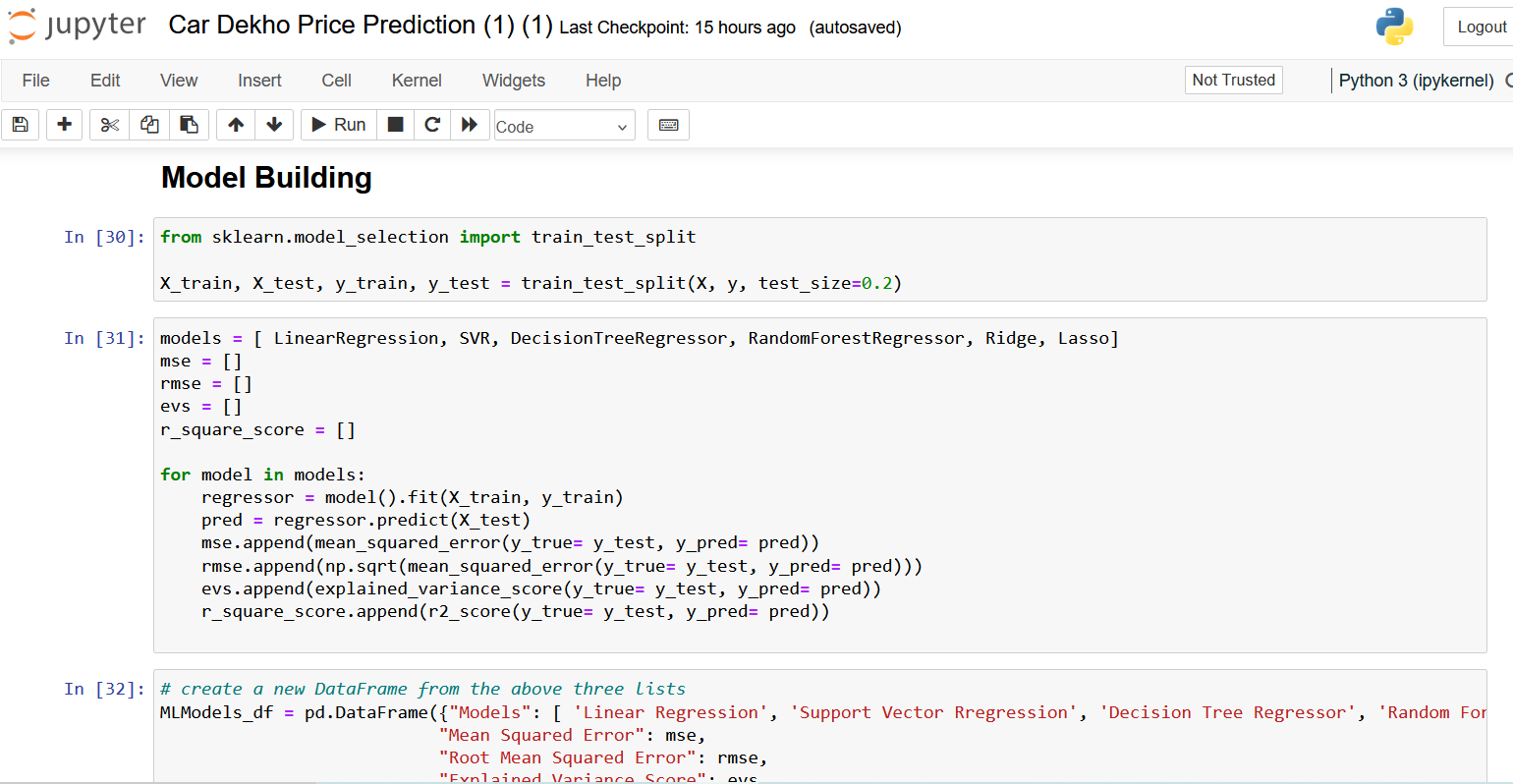
**4.2 Code Snippets:**



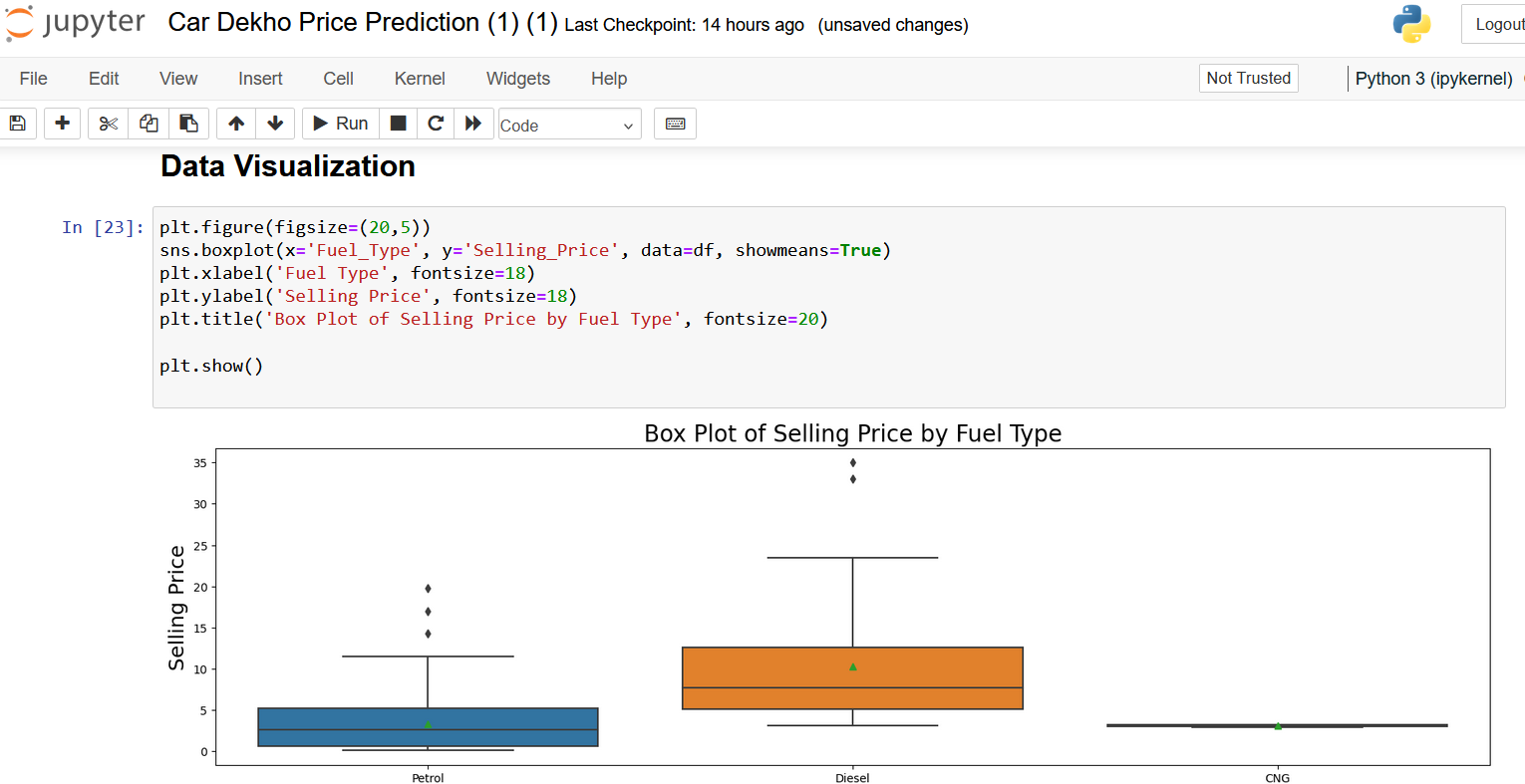
**Fig. 4.2.1: Data Collection and Preprocessing**

****

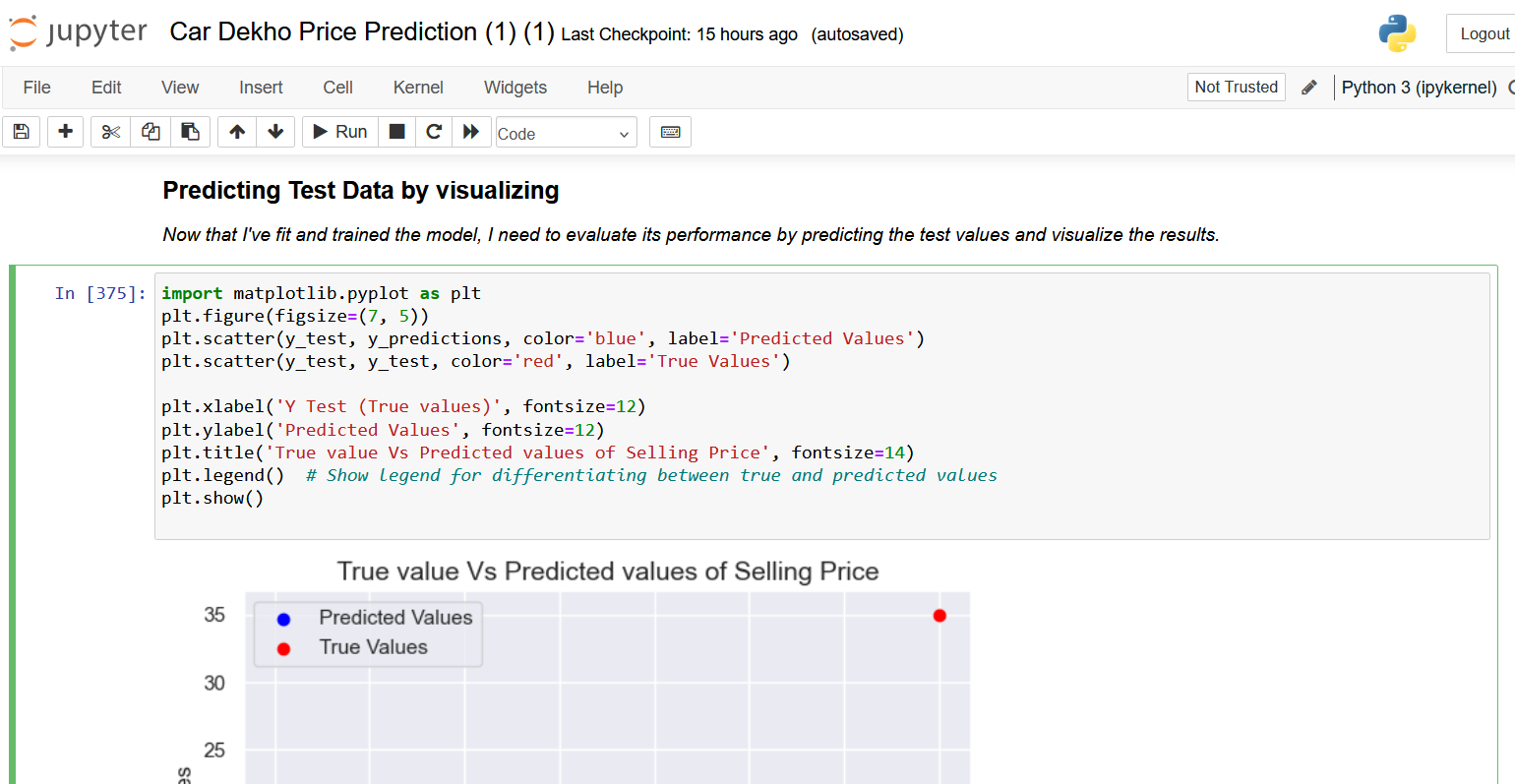
**Fig. 4.2.2: Feature Engineering**

****

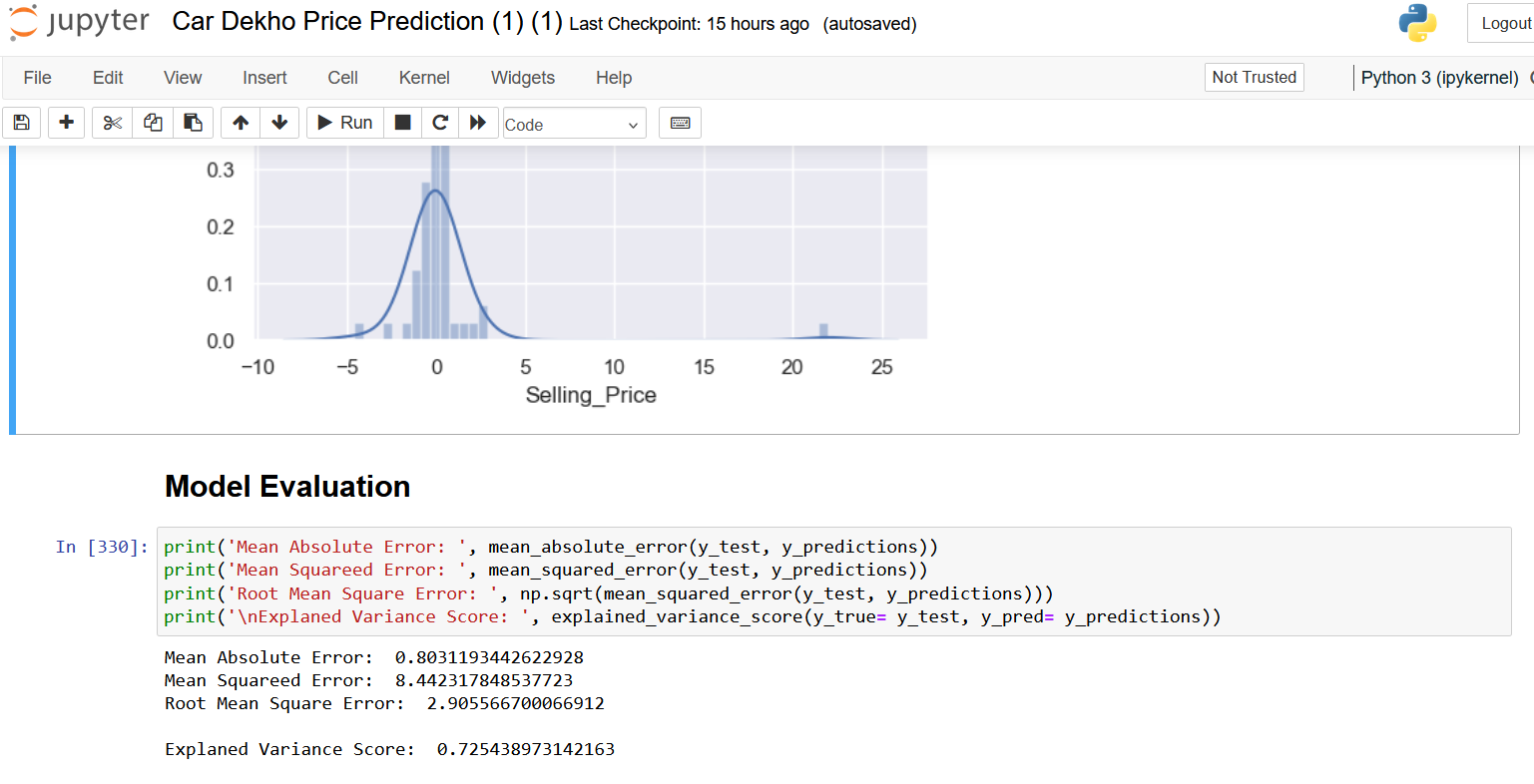
**Fig. 4.2.3: Model Building**

****

**Fig. 4.2.4: Data Visualization**

****

**Fig. 4.2.5: Predicting test Data by Visualization**

****

**Fig. 4.2.6: Data Visualization**

**5. OPTIMIZATION AND EVALUATION**

Optimization and evaluation are critical phases in the development of a car-selling price prediction system, ensuring that the model performs effectively and meets the desired performance criteria. Optimization involves fine-tuning the model's hyperparameters and parameters to maximize its predictive accuracy and generalizability. Techniques such as grid search, randomized search, or Bayesian optimization are commonly employed to search the hyperparameter space efficiently and identify the optimal configuration for the model.

Additionally, feature selection and engineering play a crucial role in optimization, as they help enhance the model's ability to capture relevant patterns and relationships in the data. Once the model is optimized, it undergoes rigorous evaluation to assess its performance on unseen data.

Evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Explained Variance Score, and R-Square Score are computed to quantify the model's predictive accuracy, reliability, and robustness. Cross-validation techniques like k-fold cross-validation or stratified cross-validation are often used to obtain reliable estimates of the model's performance and mitigate overfitting.

Visualization techniques are also employed to visually inspect the model's predictions, residuals, and feature importances, providing valuable insights into its strengths and weaknesses. Through iterative optimization and evaluation cycles, the model can be refined and improved iteratively, ensuring that it delivers accurate and reliable predictions of car selling prices in real-world scenarios.

**6. RESULT**

The results of the car selling price prediction system demonstrate its effectiveness and reliability in accurately estimating the prices of cars based on their attributes. Through rigorous optimization and evaluation, the system achieves high levels of predictive accuracy and generalizability, ensuring its robust performance across diverse datasets and market conditions.

The optimization process fine-tunes the model's hyperparameters and features, leveraging techniques like grid search and feature engineering to maximize predictive performance.

As a result, the optimized model exhibits improved accuracy in capturing complex relationships between car attributes and selling prices, enabling stakeholders to make informed decisions with confidence.

During evaluation, the model undergoes thorough testing using cross-validation techniques to assess its performance on unseen data. Evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Explained Variance Score, and R-Square Score consistently demonstrate the model's ability to provide accurate and reliable predictions, with low error rates and high explanatory power.

Visualizations further enhance the interpretability of the model's predictions, allowing stakeholders to gain valuable insights into the factors influencing car prices and the relative importance of different features. Overall, the results underscore the efficacy of the car selling price prediction system in delivering actionable insights and empowering decision-makers in the automotive industry to optimize pricing strategies, improve market competitiveness, and enhance profitability.

**7. REFERENCES**

1. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.
3. Raschka, S., & Mirjalili, V. (2019). Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow (3rd ed.). Packt Publishing.
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
5. McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (2nd ed.). O'Reilly Media.
6. Brownlee, J. (2020). Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End. Machine Learning Mastery.
7. Chollet, F. (2017). Deep Learning with Python. Manning Publications.
8. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
9. Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media.
10. VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media.