**CAR PRICE PREDICTION**

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

**Bachelor of Technology**

In

**Computer Science and Engineering**

**School of Engineering and Sciences**

Submitted by

N.MADHAV (AP21110011260),

P.GURU KIRAN (AP21110010707),

SAI KIRAN (AP21110010698).

**A picture containing text

Description automatically generated**

Under the Guidance of

**(Dr. M Krishna Siva Prasad)**

**SRM University–AP**

**Neerukonda, Mangalagiri, Guntur**

**Andhra Pradesh – 522 240**

**April, 2024**

**Certificate**

Date: 30-April-24

This is to certify that the work present in this Project entitled “**CAR PRIZE PREDICTION”**has been carried out by **[N.Madhav, P.Guru Kiran, M.Sai Kiran]** under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology/Master of Technology in **School of Engineering and Sciences**.

Dr. M Krishna Siva Prasad

**Supervisor**

**Acknowledgement**

We extend our deepest gratitude to Dr Krishna Siva Prasad for his valuable guidance, unwavering support, and mentorship throughout the course of our Applied Data Science (ADS). Dr Siva Prasad's expertise, encouragement, and dedication have been instrumental in shaping the success of this project endeavour.

His insightful feedback, constructive criticism, and commitment to academic excellence have significantly contributed to the development of this project.

We express our sincere thanks to Dr Krishna Siva Prasad for his mentorship, which has been a guiding force in our academic journey.

Yours Sincerely,

N.MADHAV (AP21110011260),

P.GURU KIRAN (AP21110010707),

SAI KIRAN (AP21110010698).

# Contributions:

**Madhav - Data Preparation and Preprocessing**

Contribution: Madhav takes the lead on data cleaning, preprocessing, and feature engineering. He handles missing values, outliers, and normalization of data to ensure the dataset is ready for modelling. Madhav also conducts initial exploratory data analysis to identify patterns, correlations, and potentially useful features that could enhance model performance.

**Guru Kiran - Model Development and Optimization**

Contribution: Guru Kiran is responsible for building and tuning various machine learning models. His experiments with different algorithms, including both traditional approaches like Logistic Regression and more complex ones like Random Forest and decision tree. Guru Kiran focuses on adjusting hyperparameters, implementing cross-validation, and utilizing techniques like grid search to find the optimal settings for each model.

**Sai Kiran– Analysis and Validation**

Contribution: Sai Kiran takes charge of evaluating the models using appropriate metrics such as accuracy, precision, recall, F1-score, and ROC curves. He analyses the outcomes to determine which models perform best and interprets the results to provide insights into the predictive power and reliability of each. Sai Kiran also performs statistical tests and validations to ensure the robustness of the findings.

**Collective Contribution:** Together, the team works collaboratively to ensure that every aspect of the project from data handling to final presentation is thorough and polished. We regularly meet to discuss our progress, troubleshoot issues, and ensure that the project aligns with the course objectives and grading criteria. Our collective efforts lead to a comprehensive understanding of how different models can be applied to predict credit card defaults, showcasing our ability to apply theoretical knowledge to a practical problem, which is essential in data science education.

# Novelity:

Diverse Model Deployment: multiple algorithms such as Linear regression, model evalution matrix, Decision Trees, KNN, Random Forest each optimized for the dataset, ensures a robust predictive framework.

Advanced Metrics Evaluation: The project goes beyond standard accuracy measures, using precision, recall, F1 score, and ROC-AUC to provide a nuanced assessment of model performance.

Focus on Tuned Performance: Each model is rigorously tuned to enhance its predictive power and ability to generalize, critical for applications in the dynamic financial sector.

Predictive Power of Socioeconomic Factors: The analysis incorporates demographic insights, enhancing the understanding of how different factors influence credit behaviour and facilitating tailored risk management strategies.

Real-time Application Potential: Emphasis on real-time capable models like SGD aligns with the needs of financial institutions for immediate risk assessments.

Ethical and Regulatory Compliance: The project addresses potential biases and ensures compliance with regulatory standards, promoting ethical AI usage in financial practices.

Integration with Business Processes: Predictive insights are designed to integrate seamlessly into business workflows, enhancing decision-making in car management.

# Abstract:

The automotive industry is continuously evolving, with fluctuating market dynamics challenging both buyers and sellers in making informed decisions about car prices. In response, data science methodologies offer a promising avenue for predictive analytics. In this project, titled "Car Price Prediction," we employ a diverse set of machine learning techniques, including Linear Regression, Decision Trees, K-Nearest Neighbors (KNN), and Random Forest, to forecast car prices with enhanced accuracy.

Through comprehensive data collection and preprocessing, we assemble a rich dataset comprising various attributes such as vehicle specifications, market trends, and historical sales records. Leveraging this data, we implement each of the aforementioned machine learning algorithms to construct predictive models capable of estimating car prices across different segments and regions.

To assess the performance of our models, we employ a range of model evaluation metrics tailored to each algorithm. These metrics include Precision-Recall Curve, ROC Curve, F1 Score, R-squared (R2) Score, Mean Absolute Error (MAE), Mean Squared Error (MSE), Gini Impurity, Information Gain, Elbow Method, Out-of-Bag Error, and more. Through rigorous evaluation and comparison, we identify the strengths and weaknesses of each technique in the context of car price prediction.

Our results demonstrate the effectiveness of ensemble methods such as Random Forest in handling complex relationships within the data, while also highlighting the interpretability of linear models like Linear Regression. Additionally, Decision Trees provide valuable insights into feature importance, aiding in understanding the driving factors behind car pricing.

# Introduction:

In the rapidly evolving landscape of automotive markets, understanding the intricacies of car pricing is paramount for both buyers and sellers. Amidst this complexity arises a pioneering data science project: Car Price Prediction. Leveraging a diverse array of machine learning techniques, including Linear Regression, Decision Trees, K-Nearest Neighbors (KNN), and Random Forest, this project endeavors to unlock the secrets of car valuation and provide actionable insights to stakeholders.

At its core, Car Price Prediction seeks to address the perennial challenge of accurately forecasting the prices of automobiles. By harnessing the power of advanced analytics and predictive modeling, the project aims to empower consumers with the knowledge to make informed purchasing decisions and assist sellers in setting competitive prices.

The utilization of multiple machine learning algorithms underscores the project's commitment to comprehensively exploring the nuances of car pricing dynamics. From the linear relationships captured by Linear Regression to the intricate decision-making of Decision Trees and the ensemble learning capabilities of Random Forest, each technique offers unique perspectives on the underlying patterns driving car prices.

Moreover, the inclusion of K-Nearest Neighbors adds further depth to the predictive capabilities of the project. By leveraging sophisticated mathematical principles and proximity-based algorithms, these methods enhance the accuracy and robustness of price predictions, catering to the diverse needs of users across different market segments.

Central to the success of Car Price Prediction is the meticulous evaluation of model performance through a comprehensive set of evaluation metrics. By employing techniques such as precision-recall curves, ROC curves, R-squared scores, and feature importance analyses, the project aims to provide actionable insights into the strengths and limitations of each predictive model.

# Background:

In the realm of automotive sales and purchasing, pricing has always been a crucial factor influencing decision-making processes. Traditional methods of determining car prices often relied on subjective assessments, market trends, and manual analysis, leading to inconsistencies and inefficiencies in pricing strategies. However, with the advent of data science and machine learning techniques, a new era of predictive analytics has emerged, offering unprecedented insights into pricing dynamics.

The "Car Price Prediction" project is born out of the necessity to leverage advanced data science methodologies to accurately forecast the prices of automobiles. By harnessing the power of various machine learning algorithms, including Linear Regression, Decision Trees, K-Nearest Neighbors (KNN), and Random Forest, the project aims to revolutionize the way car prices are determined.

Data is the cornerstone of this endeavor. The project aggregates vast amounts of data from diverse sources, including historical sales records, market trends, vehicle specifications, economic indicators, and demographic information. This rich dataset serves as the foundation for training and validating the machine learning models.

# Description:

Data Collection and Preprocessing:

The project begins with the collection of a comprehensive dataset containing various attributes of cars, such as make, model, year, mileage, fuel type, engine size, and more.

Data preprocessing involves cleaning the dataset, handling missing values, encoding categorical variables, and scaling numerical features to ensure compatibility with machine learning algorithms.

Model Selection and Training:

We employ several machine learning algorithms to build predictive models for car price estimation: Linear Regression, Decision Trees, K-Nearest Neighbors (KNN), and Random Forest.

Model Evaluation and Metrics:

For each trained model, we employ a comprehensive set of evaluation metrics to assess its performance:

Decision Trees and Random Forest models are evaluated based on criteria like Gini Impurity, Information Gain, and feature importance.

KNN performance is assessed using metrics like misclassification rate, distance metrics evaluation, and cross-validation.

Additionally, we visualize performance metrics using precision-recall curves, ROC curves, and confusion matrices to gain deeper insights into model behavior and decision-making.

Model Comparison and Selection:

After evaluating the performance of each model, we compare their strengths and weaknesses based on the defined metrics.

We select the most accurate and reliable model for car price prediction based on its performance across multiple evaluation criteria.

Deployment and Future Work:

Once the optimal model is identified, we deploy it into a user-friendly interface or integrate it into an existing platform for real-time car price prediction.

Continuous monitoring and refinement of the model will be carried out to adapt to changes in the automotive market and improve prediction accuracy over time.

# Proposed solution using Data Science Technique:

Based on the analysis of various machine learning models using ROC curve and accuracy metrics, we have identified decision tree and stochastic gradient descent as the top-performing models for our car prize project. Therefore, our proposed solution for addressing the challenges in the credit card industry in Taiwan is as follows:

# Data Collection and Preprocessing:

# Gather a comprehensive dataset containing information about car listings, including features like make, model, mileage, year, condition, location, and price.

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

# Encode categorical variables using techniques like one-hot encoding or label encoding.

# Scale numerical features to ensure they have similar magnitudes, improving model performance.

# . Feature Engineering:

# Extract additional features such as car age, engine displacement, and fuel type from existing variables.

# Explore interactions between features and create new ones if necessary.

# Use domain knowledge to identify relevant features that could enhance model performance.

# . Model Selection:

# Experiment with multiple machine learning algorithms, including Linear Regression, Decision Trees, K-Nearest Neighbors (KNN), and Random Forest.

# Train each model using the preprocessed dataset and evaluate their performance using appropriate evaluation metrics such as RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and R-squared.

# Select the top-performing models based on evaluation metrics for further optimization.

# . Model Evaluation and Tuning:

# Perform cross-validation to assess the generalization performance of the models and mitigate overfitting.

# Fine-tune hyperparameters for each selected model using techniques like grid search or randomized search.

# Validate model assumptions and ensure they hold true for the dataset.

# Compare the performance of different models using advanced evaluation metrics like precision-recall curves, ROC curves, and feature importance scores.

# DATA VISUALISATION:

# 

# 

# Conclusion:

In the realm of machine learning and data science, Linear Regression, Model Evaluation Metrics, Decision Trees, K-Nearest Neighbors (KNN), and Random Forest algorithms stand as pillars of predictive analytics, each with its unique strengths and applications.

As we journey through the landscape of machine learning algorithms, we find that each tool in our arsenal offers a unique perspective and set of capabilities. Whether it's the elegance of Support Vector Machines, the simplicity of Linear Regression, or the versatility of Random Forests, these algorithms empower us to extract insights from data and unlock the mysteries of the world around us. In the ever-evolving field of data science, their contributions continue to shape the way we understand and interact with the world.

# References:

https://www.geeksforgeeks.org/machine-learning/

https://in.mathworks.com/discovery/machine-learning-models.html