Saudi Arabia Used Cars' Price Prediction using Machine Learning

Rama Hassan[1], Lubna Alnajran[2], Raghad Basamh[3] , Sara Alotaibi [4]

[1234]*CCIS Department, Prince Sultan University, Riyadh, Saudi Arabia*

[1]222410590@psu.edu.sa

[2]222410296@psu.edu.sa

[3]222410359@psu.edu.sa

[4]222410889@psu.edu.sa

***Abstract*— This project presents a machine learning model designed to accurately predict used car prices specifically in the Saudi market given a set of features. The main objective is to help create a fair and transparent used car market by offering an unbiased data driven model. We used a dataset of real used car vehicle listings from the Saudi Arabian market containing several features with a variety of brands and conditions. Using the dataset, we performed data cleaning, exploratory data analysis, and feature engineering to achieve relevant predictor features such as car make, mileage, year, etc.**

The dataset was preprocessed using dummy encoding to convert categorical features into numerical representations in order to enable the training of our prediction model. Multiple models were implemented and evaluated, such as Linear Regression (LR), Random Forest (RF), Gradient Boosting, K-Nearest Neighbors (KNN), and a hybrid ensemble. Random Forest achieved the strongest performance ($R^2 = 0.80$) with the lowest mean absolute error, and therefore was selected for deployment as an interactive Streamlit application.

The deployed system allows users to input real car attributes and receive automated price predictions. The results revealed a strong performance for recurring car brands within the dataset, with limitations for rare luxury makes due to their rarity in the dataset. The model demonstrates the development of a sustainable and user-friendly valuation tool aligned with responsible consumption practices.

***Keywords*—– Saudi Used Cars, Linear Regression, Random Forest, Car Price Prediction, Gradient Boosting**

I. Introduction

The used car market in Saudi is one of the largest automotive markets in the region, with thousands of listings being posted daily on various resale platforms, but despite its abundance, it suffers from inconsistent pricing and a lack of standardized valuation methods resulting in high variability. This results in buyers often relying on subjective price judgements, personal experience and generally facing difficulty determining a fair market value. Such variability increases the risk of overpricing, undervaluation, and uninformed financial decisions.

Our project utilizes machine learning by developing a data-driven, automated valuation model capable of analyzing a large number of records and learning pricing patterns to achieve accurate used car predictions. This was implemented using real historical data, the dataset used contained a variety of 8035 vehicles with their essential attributes such as Make, Year, Color, Options package, Engine Size, Gear Type, and Mileage. We began by performing data cleaning and exploratory data analysis (EDA), followed by feature engineering which involved the conversion of categorical attributes into dummy integer variables to enable model training. To achieve the most accurate and reliable predictions, we trained multiple regression models, including a hybrid model, to assess each model's accuracy and select the best performing algorithm. Finally, we deployed the prediction model on Streamlit to offer an accessible and reliable prediction interface to consumers, where users are able to input certain vehicle features and obtain price estimates.

Our project's objective is to revolutionize the car resale industry by deploying a consumer-friendly system that predicts car prices based on real historical data, offering customers fair estimates based on quantifiable features, resulting in improved price transparency, and increase in consumer protection and a standardized market.

II. Dataset Description

The used dataset, which was sourced from Kaggle, consists of real used car listings collected from the Saudi market. It includes a mix of numerical, categorical, and descriptive features, allowing for both statistical analysis and machine learning based predictions. Each vehicle had the following features influencing its market value:

TABLE I
USED DATASET COLUMNS

Columns	Explanation
Make	Which references the car manufacturer (e.g., Toyota, Hyundai, Ford, etc.), 59 different makes were present in the dataset.
Type	The make's model.
Year	Which year was the car manufactured
Origin	Which origin in Saudi are the cars sold in.
Color	Color of the car
Options	Refers to the equipment level. Three options were present: Standard, Semi-Full, and Full
Engine_size	Refers to the engine's fluid capacity in liters.
Fuel_Type	Gasoline, Diesel, Hybrid.
Gear_Type	Can be either Automatic or Manual.
Mileage	The number of kilometres travelled or covered.
Region	Defines the region the car is listed from
Price	Refers to the final selling price, which is our target variable.
Negotiable	Initially indicated whether the price was negotiable or final. It was set to True for cars priced at 0 SAR.

III. METHODOLOGY

A. Data Cleaning and Preprocessing

The loaded raw dataset required to be cleaned in order to achieve a reliable predictive model. Approximately 25% of used cars were listed at a price of 0 SAR, which is unrealistic and therefore removed. Since the filtering process involved applying constraints that restrict the price from being set to a number under 2,000 SAR, the Negotiable column has since proven redundant and no longer needed since it was only set to True when the price was 0 SAR. The dataset contained unrealistic values where vehicle prices started from 0 SAR and mileage values exceeded 1,000,000 KM, which resulted in the elimination of such outliers. In addition, the removal of duplicated rows was implemented and column Type was renamed to Model.

B. Exploratory Data Analysis (EDA)

Exploratory data analysis was performed to reveal relationships that might have been overlooked and learn the trends occurring between variables. As shown in Fig. 1, Fig 2., Fig 3., and Fig.,4.

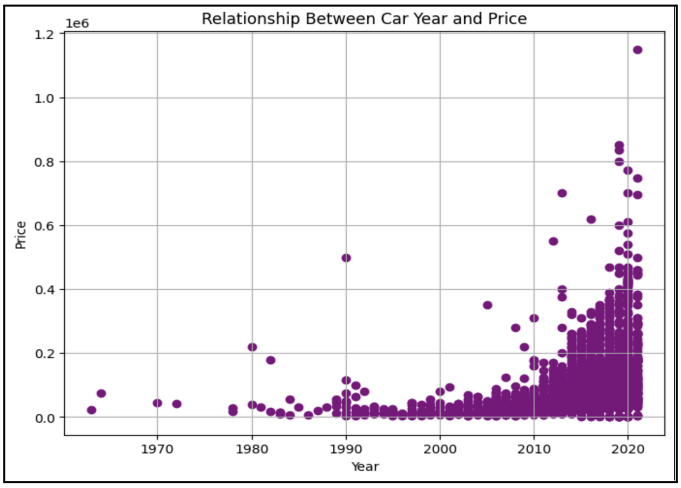


Fig. 1

Scatterplot of Price and Year

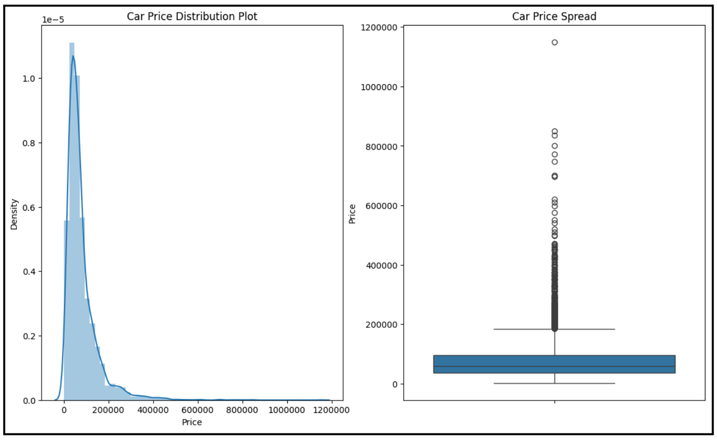


Fig. 2

Price Distribution with a significant difference between the mean and the median of the price distribution indicating data points are far spread out from the mean, which indicates a high variance in the car prices.

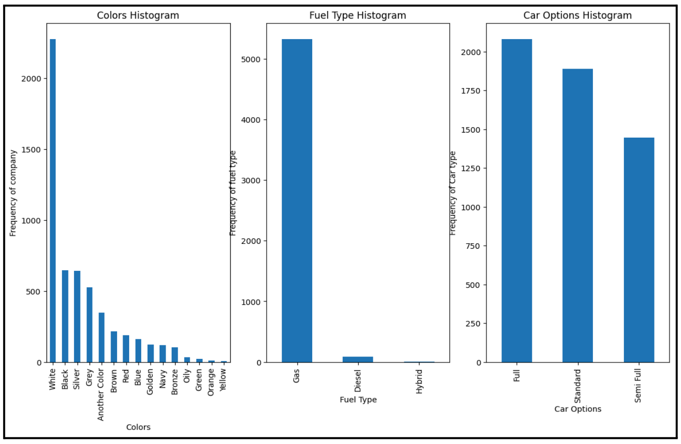


Fig. 3 Feature Frequency Plots

The histograms reveal that white cars were the most frequently listed, saturating the market by a high margin, fuel type is most commonly gasoline, and full option cars are the most frequently listed.

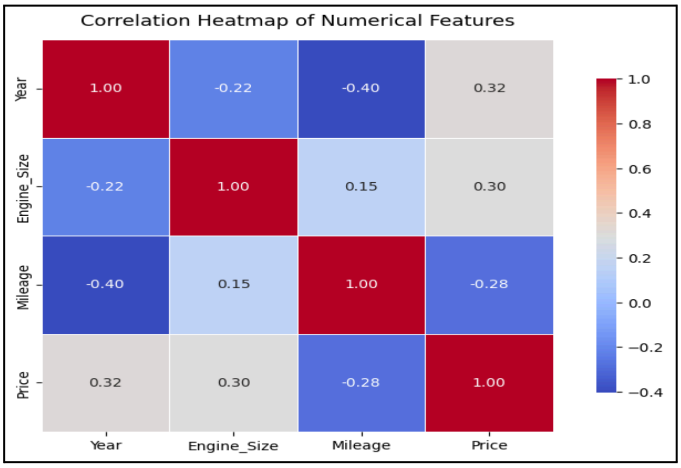


Fig. 4 Correlation Heatmap

The correlation heatmap of the numeric features reveals a negative relationship between Year and Mileage, showing that older cars tend to have higher mileage. In addition, Mileage has a negative correlation with Price, suggesting that heavily used cars are generally sold for lower prices. Alternatively, Year and Engine_Size both have positive relationships with Price, indicating that newer vehicles with larger engines tend to be more expensive.

C. Feature Engineering

Feature engineering was an integral step for building the model, since converting raw categorical values into numerical representations was essential for the model to be able to learn patterns. The following steps were taken: Calculating Car Age by subtracting the car's year by the year 2025, and converting significant categorical features to integers using dummy variables (0/1 columns) to enable machine learning model training, then they were stored in a new variable "dummies" to be put into the dataframe. Finally, the initial categorical values were dropped. As for the model development, the target column Price was dropped from the features and was assigned in the output variable, the data was split into training and testing, where 70% of the data was used for training and the rest of the 30% of the data was used for testing. Then the following supervised models were trained and evaluated to identify the best performing model to deploy:

1) Linear Regression : Linear regression was used as a baseline model. The model gave us a reference point for comparing more advanced models. It assumes a linear relationship between the car features (e.g., model, year, mileage) and the target price such that the target variable *y* is a linear function of the features *X*. It was simple, fast, and easily interpretable yet it performed poorly since the used-car market in Saudi is not linear; sellers set prices differently, regions vary, and car options/features don't increase price in a straight line. As a result, Linear Regression struggled to capture the complexity of actual pricing.

2) Random Forest Regressor*:* To increase accuracy and robustness, Random Forest constructs several decision trees and averages the forecasts they make, involves bootstrapping multiple subsets from the training data, training a decision tree on each subset, and averaging the predictions from all trees for regression tasks. This allows it to capture complex patterns without significant overfitting. Random Forest performed significantly better with an 80% accuracy rate than Linear Regression and became one of our primary models as it was able to handle non-linear relationships and deal with noisy data outliers much better. This model suits the Saudi used-car market because the pricing in Saudi is inconsistent, and depends heavily on hidden patterns. RF found those patterns without requiring heavy assumptions.

3) Gradient Boosting Regressor*:* An additive model by sequentially training shallow trees, where each tree corrects the residuals of the previous iteration. This model was used as it captures subtle pricing patterns associated with brand reputation, region, model popularity, and additional options.

The model delivered high performance with an accuracy rate of 80% but was more sensitive to overfitting and required careful hyperparameter tuning.

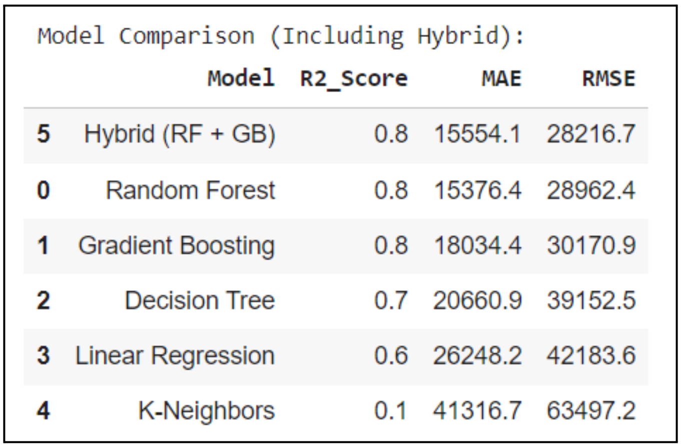
4) Decision Tree Regressor: This algorithm partitions the feature space using hierarchical splits that minimize variance within each node. It emerged as one of the foundational models in our project, leveraged to uncover the multifaceted factors influencing car prices with an accuracy rate of 70%.

5) K–Nearest Neighbors (KNN): KNN estimates price by averaging the prices of the closest neighbors in feature space based on Euclidean distance. The use of this model was due to used car buyers often relying on comparisons between similar vehicles, making KNN conceptually relevant. Yet, this model's performance degraded significantly due to sensitivity to scaling, high dimensionality, and irregular pricing with an alarming accuracy score of 10%.

6) Hybrid Ensemble (Random Forest and Gradient Boosting): A hybrid model was constructed for innovation purposes and done by averaging the predictions of the two best performing algorithms. The rationale behind this was that ensemble averaging reduces model-specific biases and helps stabilize predictions by combining the strengths of bagging variance reduction and boosting error correction. Producing smoother and more robust price estimates, the hybrid model achieved the most stable performance across evaluation metrics.

D. Model Performance Evaluation

Each model was evaluated using three metrics widely used in regression analysis. The coefficient of determination (R^2) measures how much variance in price is explained, mean absolute error (MAE) measures average absolute average deviation, root mean squared error (RMSE) penalizes large errors more heavily.



Model Comparison (Including Hybrid):

	Model	R2_Score	MAE	RMSE
5	Hybrid (RF + GB)	0.8	15554.1	28216.7
0	Random Forest	0.8	15376.4	28962.4
1	Gradient Boosting	0.8	18034.4	30170.9
2	Decision Tree	0.7	20660.9	39152.5
3	Linear Regression	0.6	26248.2	42183.6
4	K-Neighbors	0.1	41316.7	63497.2

E. Discussion of Results

The results demonstrate that tree-based ensemble models outperform linear and distance-based approaches, confirming that used car pricing in Saudi Arabia follows a complex, non-linear structure influenced by interactions between features. Random Forest (RF) and Gradient Boosting (GF) achieved the strongest individual performance, each reaching $R^2 = 0.8$, indicating strong explanatory power. The hybrid ensemble improved stability and slightly reduced prediction error by combining the variance reduction of Random Forest with the error-correcting capability of Gradient Boosting. Decision Tree performed moderately but suffered from overfitting, consistent with expectations, Linear Regression showed limited predictive capability, verifying that price does not follow a linear trend. KNN produced the lowest accuracy due to sensitivity to high-dimensional data and inconsistent pricing patterns. Overall, the hybrid ensemble provides the highest quality and most reliable predictions and it was strictly added for innovation purposes and was not used for our application deployment. Random Forest was pursued for the rest of our project and chosen as a candidate for our application deployment as shown in Fig. 5. and Fig.6. ,

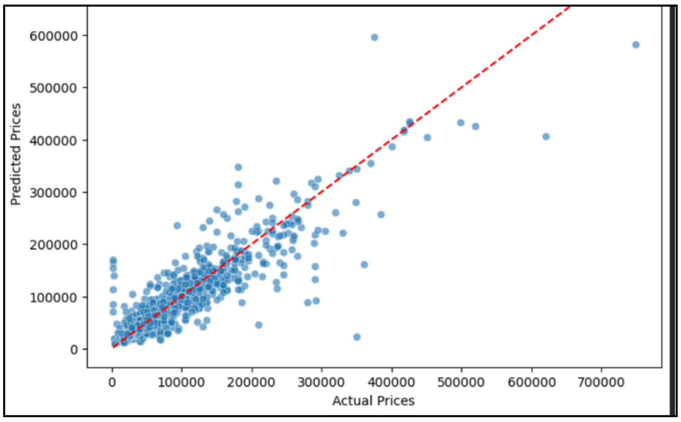


Fig. 5 Scatter Plot of Actual Price vs. Predicted Price

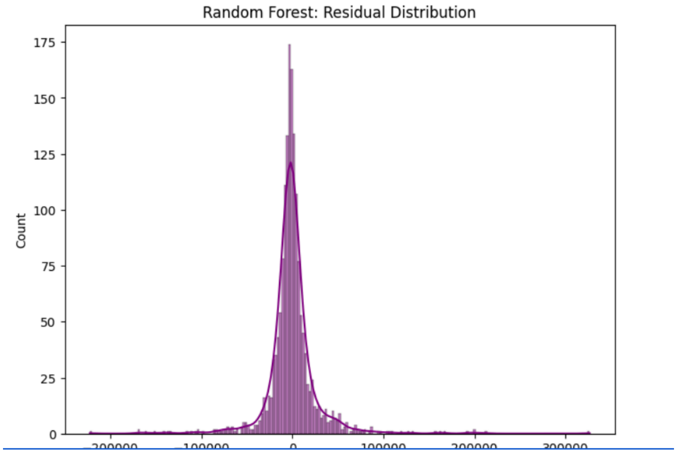


Fig. 6 Histogram Plot of Residual Distribution

IV. Conclusion

In this project, we developed a complete machine learning pipeline for predicting used-car prices in the Saudi Arabian market using real-world data. After extensive data cleaning, by removing unrealistic entries, filtering noise, and converting categorical attributes into dummy variables, we trained and evaluated several regression models to determine the most accurate predictor. Our experiments showed that the Random Forest Regressor achieved the best performance with an R^2 score of 0.80, making it highly reliable for capturing the complex, non-linear relationships between car features and market value.

The exploratory data analysis revealed clear market trends, including the influence of mileage, year, engine size, and brand on car pricing. These insights were consistent with real consumer behavior and validated our modeling decisions. Furthermore, the deployment of the model through a fully interactive Streamlit web application makes the solution practical, accessible, and instantly usable for real-world decision-making.

On a broader scale, this project demonstrates how data-driven pricing can bring fairness, transparency, and consistency to the used-car market. By providing unbiased, model-generated estimates, our tool helps both buyers and sellers make more informed decisions, ultimately contributing to a more balanced and trustworthy automotive ecosystem.

F. References

Conclusions

This project revealed that machine learning can successfully predict used car prices in Saudi Arabia. By analyzing main characteristics such as brand, mileage, engine size, and production year, the model provides fair, data-centric prices. These predictions encourage market transparency, help sellers set fair prices, protect buyers from overpaying, and promote sustainable consumption by increasing the lifespan of vehicles. Future integration into online car-selling platforms could regulate pricing across the country.

Acknowledgment

The team members wish to acknowledge the support of the CS316 lab tutorials from Prince Sultan University, which provided guidance on different types of modeling and Python techniques.

References

[1] Saudi Arabia Used Cars Dataset, Kaggle. [Online]. Available: https://www.kaggle.com/datasets

[2] "What is a Regression Model?" Perforce Blog. [Online]. Available: https://www.perforce.com/blog/ims/what-is-regression-model

[3] Stack Overflow, "Logistic Regression for Non-Categorical Data Prediction." [Online]. Available: https://stackoverflow.com/questions/68472062/logistic-regression-for-non-categorical-data-prediction

[4] [4] CS316 Lab Tutorials, Prince Sultan University, Riyadh, Saudi Arabia, 2025.

[5] [5] C. Bishop, Pattern Recognition and Machine Learning, New York, NY, USA: Springer, 2006.

[6] [6] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., San Francisco, CA, USA: Morgan Kaufmann, 2011.

[7] [7] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., New York, NY, USA: Springer, 2009.

[8] [8] P. Flach, Machine Learning: The Art and Science of Algorithms that Make Sense of Data, Cambridge, UK: Cambridge University Press, 2012.