USED CAR PRICE PREDICTION

USING MACHINE LEARNING

PROBLEM STATEMENT:

The used car market poses challenges for both buyers and sellers due to the lack of transparency and subjectivity in determining accurate prices for pre-owned vehicles. This information asymmetry often leads to inefficient transactions, where parties may either overpay or undersell. The need for a reliable method to estimate fair market values of used cars has become crucial in this complex landscape.

The aim of the project is to develop an accurate and reliable used car price prediction model which uses past data in order to derive a market value estimate for pre-owned vehicles. The model shall aim to provide an estimate of the resale value of a car depending on its various characteristics, e.g. make, model, kilometers driven, fuel type, transmission, owner type, mileage, engine, power, number of seats, year, and location, in order to help buyers and sellers make informed decisions.

Solution Approach

The proposed solution involves the development of a robust used car price prediction model using machine learning techniques. This design is motivated by several factors:

- Data-driven Accuracy
- Attributes
- Complex Relationship Handling
- Generalization
- Transparency

The implementation of the proposed solution has a significant impact on addressing the challenges posed by the used car market. Here's a few:

- Enhanced Efficiency
- Confidence and Empowerment
- Market Transparency
- Optimized Resource Allocation
- Business Growth

Exploratory Data Analysis

	count	mean	std	min	25%	50%	75%	max
Year	7253.0	2013.365366	3.254421	1996.00	2011.000	2014.00	2016.0000	2019.00
Kilometers_Driven	7253.0	58699.063146	84427.720583	171.00	34000.000	53416.00	73000.0000	6500000.00
Mileage	7251.0	18.141580	4.562197	0.00	15.170	18.16	21.1000	33.54
Engine	7207.0	1616.573470	595.285137	72.00	1198.000	1493.00	1968.0000	5998.00
Power	7078.0	112.765214	53.493553	34.20	75.000	94.00	138.1000	616.00
Seats	7200.0	5.280417	0.809277	2.00	5.000	5.00	5.0000	10.00
New_price	1006.0	22.779692	27.759344	3.91	7.885	11.57	26.0425	375.00
Price	6019.0	9.479468	11.187917	0.44	3.500	5.64	9.9500	160.00

Statistical Summary of Numeric Variables:

In this table, it shows that the minimum value of the mileage attribute is 0. The dataset is for used car. Therefore 0 mileage is not possible. Handling this 0 value is crucial when building the model. The 0 value was treated by adding 10 (as per the advice of the mentor ad it was added in the entire rows).

Exploratory Data Analysis

	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_price	Price
2328	BMW X5 xDrive 30d M Sport	Chennai	2017	6500000	Diesel	Automatic	First	15.97	2993.0	258.00	5.0	NaN	65.00
340	Skoda Octavia Ambition Plus 2.0 TDI AT	Kolkata	2013	775000	Diesel	Automatic	First	19.30	1968.0	141.00	5.0	NaN	7.50
1860	Volkswagen Vento Diesel Highline	Chennai	2013	720000	Diesel	Manual	First	20.54	1598.0	103.60	5.0	NaN	5.90
358	Hyundai i10 Magna 1.2	Chennai	2009	620000	Petrol	Manual	First	20.36	1197.0	78.90	5.0	NaN	2.70
2823	Volkswagen Jetta 2013-2015 2.0L TDI Highline AT	Chennai	2015	480000	Diesel	Automatic	First	16.96	1968.0	138.03	5.0	NaN	13.00
3092	Honda City i VTEC SV	Kolkata	2015	480000	Petrol	Manual	First	17.40	1497.0	117.30	5.0	NaN	5.00
4491	Hyundai i20 Magna Optional 1.2	Bangalore	2013	445000	Petrol	Manual	First	18.50	1197.0	82.90	5.0	NaN	4.45
6921	Maruti Swift Dzire Tour LDI	Jaipur	2012	350000	Diesel	Manual	First	23.40	1248.0	74.00	5.0	NaN	NaN
3649	Tata Indigo LS	Jaipur	2008	300000	Diesel	Manual	First	17.00	1405.0	70.00	5.0	NaN	1.00
1528	Toyota Innova 2.5 G (Diesel) 8 Seater BS IV	Hyderabad	2005	299322	Diesel	Manual	First	12.80	2494.0	102.00	8.0	NaN	4.00

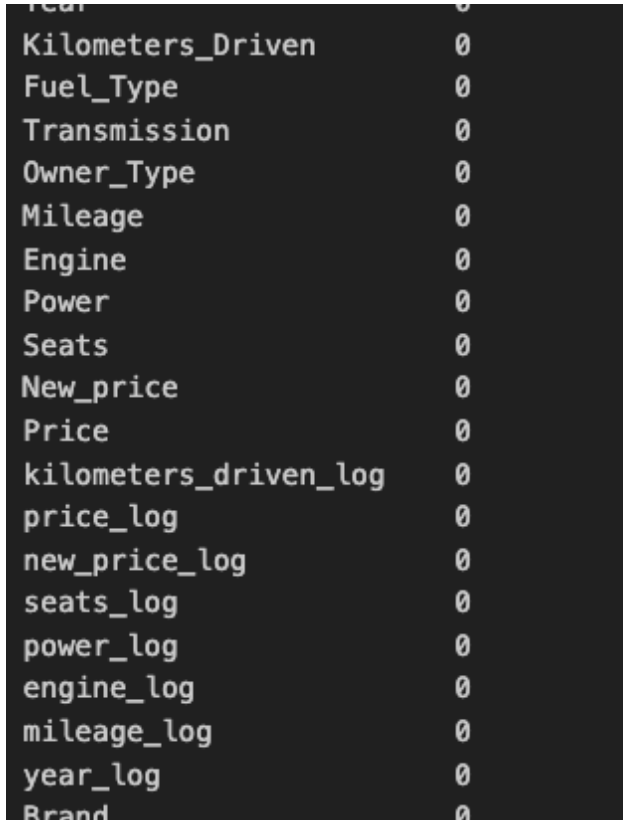
Check the Extreme Value in Kilometers_Driven Column:

As observed on the table above, row 2328 shows an extreme value, Considering that the model year of the car is only 2017, this is obviously not possible. This row was dropped prior to building the model as this will affect the performance of the model.

Handling Missing Values

Missing values were handled as part of the Feature Engineering and prior to model building.

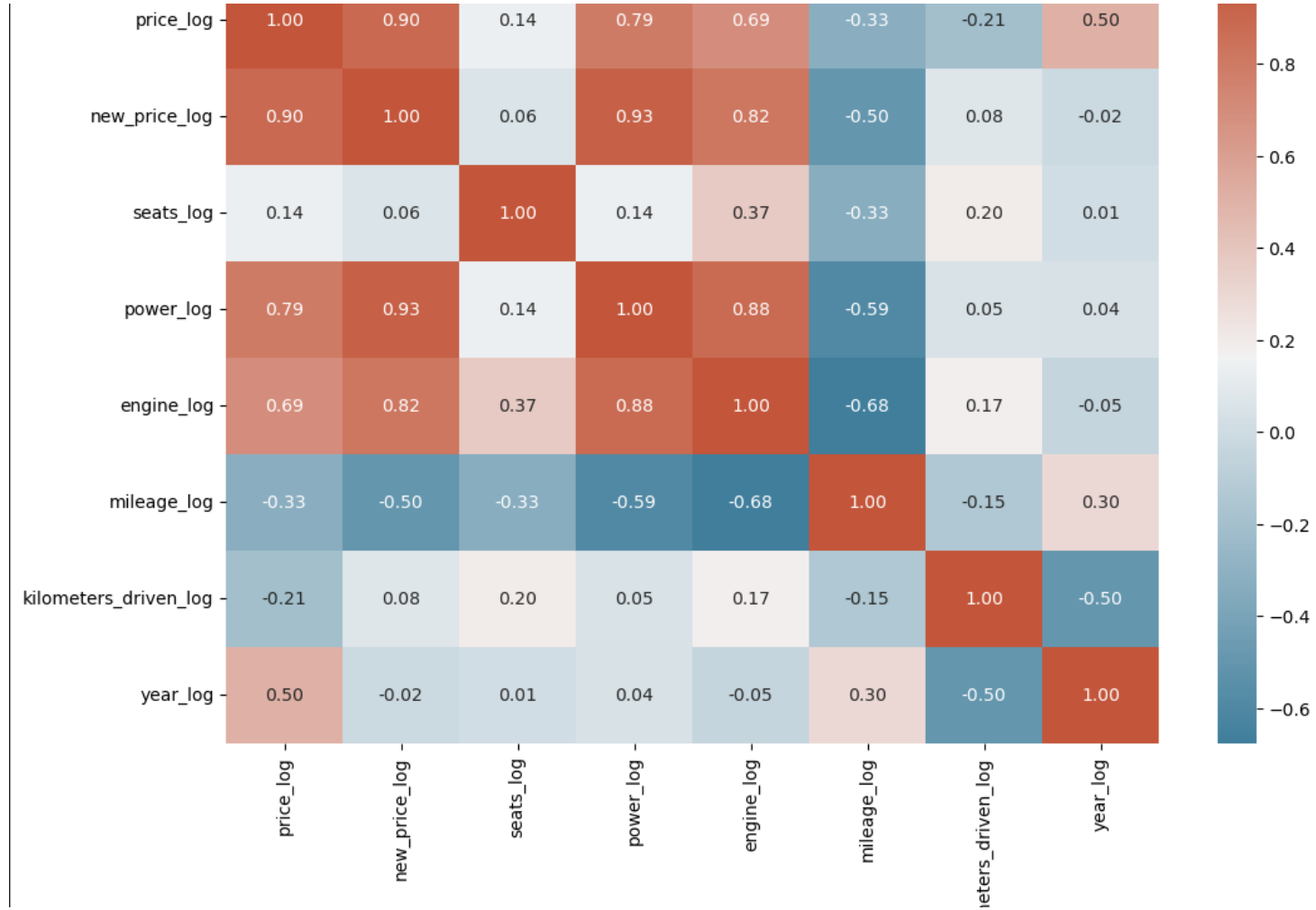
Year	0
Kilometers_Driven	0
Fuel_Type	0
Transmission	0
Owner_Type	0
Mileage	2
Engine	46
Power	175
Seats	53
New_price	6246
Price	1234
kilometers_driven_log	0
price_log	1234
new_price_log	6246
seats_log	53
power_log	175
engine_log	46
mileage_log	2
year_log	0
Brand	0



Kilometers_Driven	0
Fuel_Type	0
Transmission	0
Owner_Type	0
Mileage	0
Engine	0
Power	0
Seats	0
New_price	0
Price	0
kilometers_driven_log	0
price_log	0
new_price_log	0
seats_log	0
power_log	0
engine_log	0
mileage_log	0
year_log	0
Brand	0

2D Correlation Matrix Between Two Numerical Features

- -0.6 represents a strong negative correlation
- +1 represents a strong positive correlation
- 0 represents no correlation



Building Models

- Building Models

Feature Engineering	Model Selection	Hyperparameter Tuning	Regularization
Prior to modelling, in-depth feature engineering should be explored. This involves creating new features or transforming existing ones to capture nuanced relationships.	Experiment with a variety of regression algorithms such as Linear Regression, Random Forest, Ridge, Decision Tree.	Utilize techniques like grid search to find optimal hyperparameters for the chosen algorithms. Tuning parameters like learning rates, regularization strengths, and the tree depths can significantly impact model performance.	Apply regularization techniques like Linear Regression and Ridge to prevent overfitting and enhance model robustness.

- Suggestion to Improve the Techniques

Domain-Specific Feature Engineering	Automated Hyperparameter Tuning	Regularization Strength Grid	Feature Importance Analysis
Tailor feature engineering to incorporate domain knowledge. For instance, create features that account for model reputation, market trends, and regional variations in pricing.	Employ automated tools like Bayesian optimization to efficiently explore the hyperparameter space and discover optimal settings.	Experiment with a wider range of regularization strengths to discover the optimal balance between bias and variance.	After model training, conduct feature importance analysis to understand which attributes play a crucial role in price prediction. This can guide further feature selection and engineering efforts.

Measure of Success

- Insights:
  - In summary, the random forest model appears to be performing well in terms of both R^2 and RMSE on both training and test sets. It strikes a balance between capturing complex relationships and avoiding overfitting, making it a strong candidate for predicting the target variable in this scenario. The decision tree, while performing exceptionally well on the training set, suffers from overfitting as indicated by its poor performance on the test set.

	Linear Regression	Ridge	Decision Tree	Random Forest
R^2 on training set	0.8705	0.8706	0.9999	0.9724
R^2 on test set	0.8669	0.8644	0.7151	0.8700
RMSE on training set	3.8077	3.8061	0.0965	1.7567
RMSE on test set	3.8816	3.9180	5.6796	3.8362

Hyperparameter Tuning

- Insights:
  - In summary, both the tuned decision tree and tuned random forest models are performing well. The tuned random forest shows higher R^2 values on both training and test compared to the tuned decision tree, indicating that it captures more variance in the target variable. The RMSE values for the tuned random forest are generally smaller, suggesting that it produces more accurate predictions on average. Therefore, based on these results, the tuned random forest appears to be better-performing model in this scenario.

	Decision Tree_Tuned	Random Forest_Tuned
R^2 on training set	0.9268	0.9724
R^2 on test set	0.7857	0.8700
RMSE on training set	2.8629	1.7567
RMSE on test set	4.9257	3.8362

Proposal for the Final Solution Design

Based on the results provided above, my proposal for the final solution design is the ***Random Forest Model.***

- The random forest model demonstrates a balanced performance between training and test sets. It effectively captures complex relationships while mitigating overfitting, as indicated by its high R^2 scores on both sets.
- The random forest's ability to generalize well to new, unseen data is evident from its consistent performance across training and test sets.
- The random forest model achieves relatively low root mean squared error (RMSE) values on both training and test sets, indicating accurate predictions and small average errors.
- Handling both linear and non-linear patterns in the data, thus accommodating the complex nature of the used car market.
- The random forest model's solid performance on both training and test sets along with its ability to generalize, suggests that it is suitable for deployment in production. Its robustness and accuracy make it well-suited for providing price predictions to users.

Key Actionable for Stakeholders

- Buyers
  - Utilize the price prediction tool to make informed decisions, compare listings, and negotiate with confidence.
- Sellers
  - Set competitive yet reasonable prices for used cars, attract potential buyers, and facilitate faster transactions.
- Platform Providers
  - Integrate the price prediction model into existing platforms to enhance user engagement and differentiate services.
- Regulators
  - Promote transparency and fair trade practices within the used car industry by endorsing the adoption of the price prediction tool.

Further Analysis & Other Associated Problems

In conclusion, implementing the used car price prediction solution requires careful consideration of data, modeling, user interface, and ongoing maintenance. While it offers substantial benefits in terms of market efficiency and transparency, stakeholders must be prepared to address challenges and continuously refine the solution to ensure its continued success. Potential risks or challenges include inaccurate or incomplete data, model overfitting, market changes, and interpretability issues. Further analysis and associated problems include analyzing regional variations, developing strategies to handle outliers, and investigating the incorporation of external data sources.