CARS4U – OUR TECHNOLOGY APPROACH

Cars4U

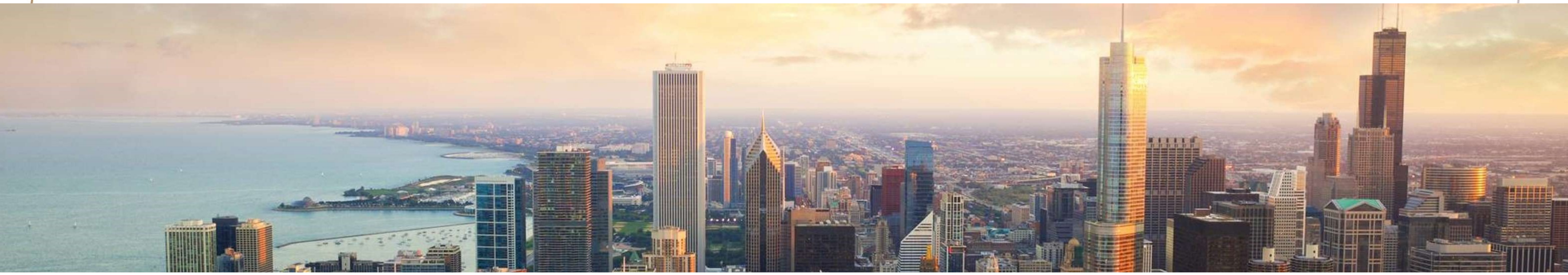
Ride with us to your next car!

AGENDA

- The Problem Definition
- Data Overview
- Modeling Process
- Next Steps
- Appendix(s)

THE PROBLEM:

WE NEED TO FIND OUR PLACE IN A GROWING MARKET.



THE USED CAR MARKET IS GROWING

- **Global market size**
  The used car market is projected to reach over **$1.5 trillion in 2027**, up from $1.2 trillion in 2020. This growth is due to changing car ownership patterns and the rise of online sales channels.

- **US market size**
  The US used car market was valued at **$191.24 billion** in 2022 and is expected to reach **$385.18 billion** by 2032.
  In 2021, over **43.1 million used light vehicles** were sold in the US, which was a slight increase from 2019. However, Cox Automotive estimated that used retail sales for 2023 were down about 3% from the previous year, due **to a limited supply of newer used vehicles.**

- Used car prices
  As of July 2024, used car prices are falling, which is good for buyers but challenging for sellers. Consumer Reports says that higher interest rates and manufacturers' focus on new models have made used cars less affordable.

https://www.coxautoinc.com/market-insights/estimated-monthly-used-vehicle-saar-and-volume/#:~:text=A%20monthly%20post%20will%20be,54%2C000%20units%2C%20from%20December%202022.

WHAT ABOUT OUR COMPETITION?

CarGurus

- Offers Unbiased Listings: CarGurus doesn't discriminate by putting paying customers' vehicles first. The best deals are listed at the top, followed by fair deals and then those that are priced over the Instant Market Value.

- Dealer Reviews: You can read CarGurus' reviews by other buyers which can tell you what to expect from a particular dealer.

- However, "Market Value" of the cars posted are abstract calculations and frequently different than other common pricing models such as the Kelly Blue Book….

TrueCar

- User-Friendly Search Tools and Vehicle Research Resources: TrueCar provides user-friendly search tools and vehicle research resources, making shopping and comparing easy

- Large Dealer Network: TrueCar has a large dealer network, which expands the inventory selection – the portal aggregates data from these dealers to obtain pricing estimates for cars.

- However, pricing data is based on averages of new cars and discounted for age – but there is little information on why a particular car may have the price that it does…

SO HOW DO WE DIFFERENTIATE?

OUR STRATEGY WILL FOCUS ON

DATA-DRIVEN INSIGHTS FOR USED-CAR PRICES AND

CLEAR PRESENTATION OF THE SPECIFIC FACTORS WHICH HAD THE GREATEST IMPACT ON THE PRICING OF THE VEHICLE –

PRESENTING ACCURATE DATA TO OUR CUSTOMERS AND INSTILLING

TRUST IN OUR PLATFORM TO BE THEIR AID IN THEIR USED CAR SHOPPING EXPERIENCE.

DATA OVERVIEW

THE BUILDING BLOCKS OF OUR PLATFORM

OUR DATA

Data Dictionary

S.No.: Serial Number

Name: Name of the car which includes Brand name and Model name

Location: The location in which the car is being sold or is available for purchase (Cities)

Year: Manufacturing year of the car

Kilometers_driven: The total kilometers driven in the car by the previous owner(s) in KM

Fuel_Type: The type of fuel used by the car (Petrol, Diesel, Electric, CNG, LPG)

Transmission: The type of transmission used by the car (Automatic / Manual)

Owner: The number of previous owners.

Mileage: The standard mileage offered by the car company in kmpl or km/kg

Engine: The displacement volume of the engine in CC

Power: The maximum power of the engine in bhp

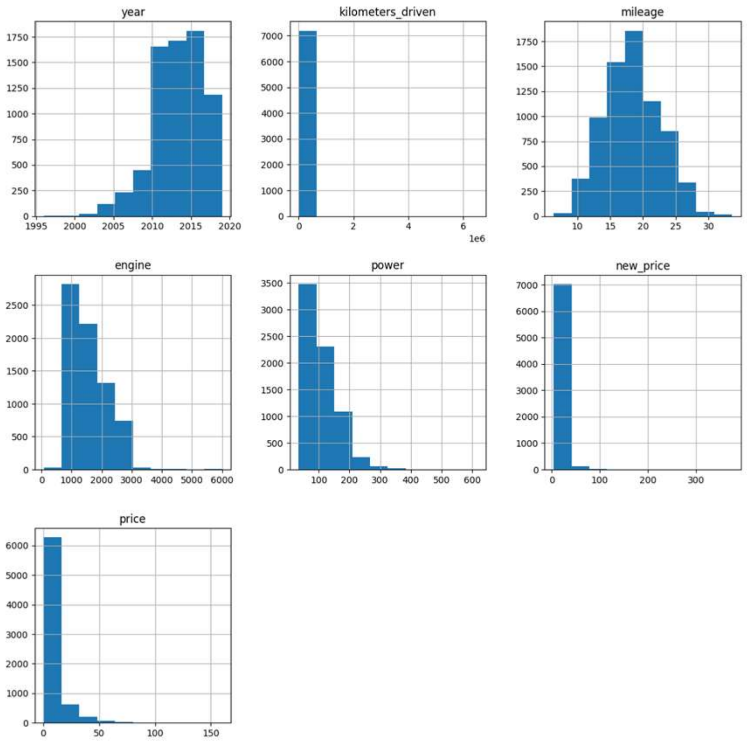
Seats: The number of seats in the car

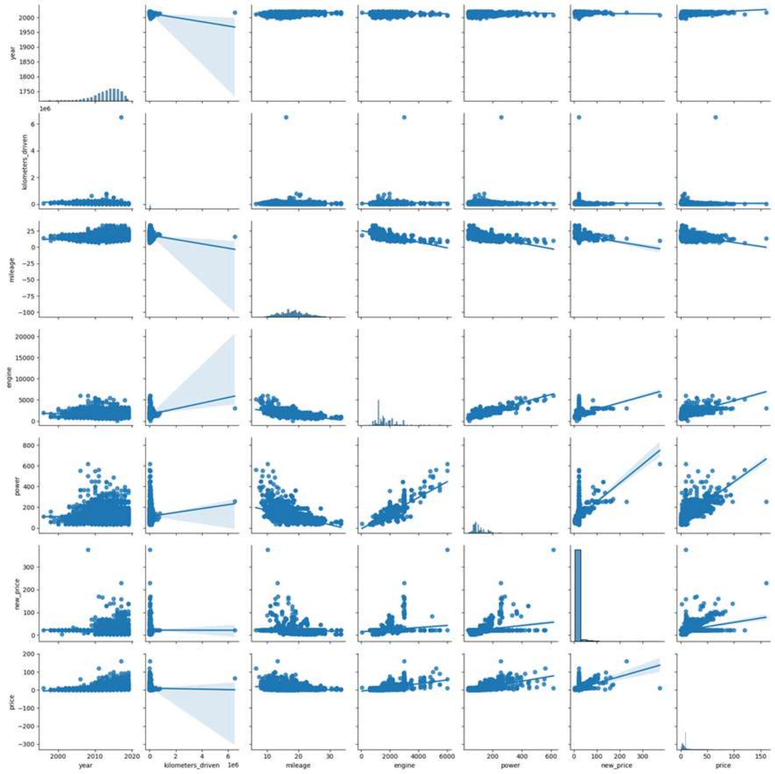
New_Price: The price of a new car of the same model in INR 100,000

Price: The price of the used car in INR 100,000 **(Target Variable)**

	S.No.	Year	Kilometers_Driven	Mileage	Engine	Power	Seats	New_price	Price
Count	7253.00	7253.00	7.25	7251.00	7207.00	7078.00	7200.00	10006.00	6019.00
Mean	3626.00	2013.00	5.87	18.14	1616.57	112.77	5.28	22.78	9.48
Std.	2093.00	3.25	8.44	4.52	595.29	53.49	0.81	27.76	11.19
Min	0.00	1996.00	1.71	0.00	72.00	34.20	2.00	3.91	0.44
25%	1813.00	2011.00	3.40	15.17	1198.00	75.00	5.00	7.89	3.50
50%	3626.00	2014.00	5.34	18.16	1493.00	94.00	5.00	11.57	5.64
75%	5439.00	2016.00	7.30	21.10	1968.00	138.10	5.00	26.04	9.95
Max	7252.00	2019.00	6.50	33.54	5998.00	616.00	10.00	375.00	160.00

UNIVARIATE AND BIVARIATE ANALYSIS

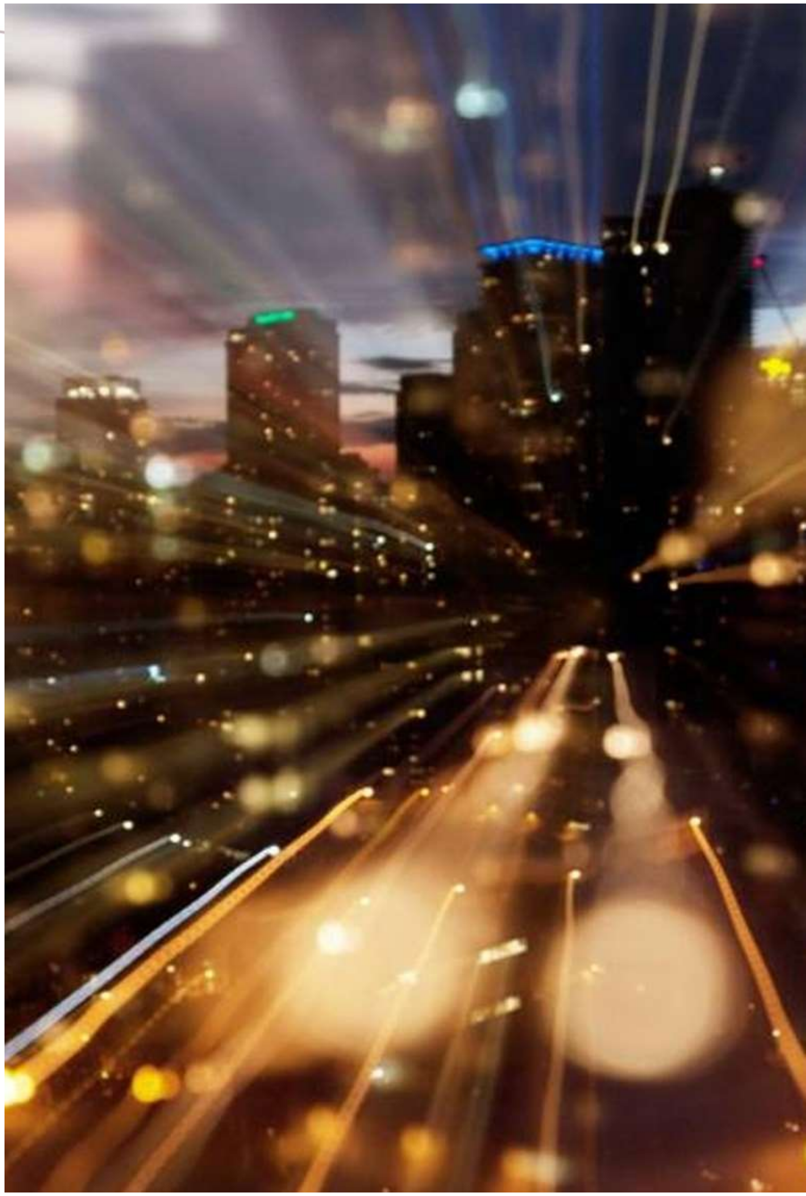




DATA SCRUBBING & CLEANUP

The following Data Scrubbing, Analytics, and Cleanup tasks were performed:

1. **Dropped Rows** where # of Seats was unknown (8 rows).
2. Performed **Mean Imputation** on other rows with empty data
3. Performed **Mean Imputation** on unrealistic Mileage (fuel economy) values
4. **Log Transformation** on Kilometers_Driven, Price, New_Price to normalize skewness.
5. **Separated vehicle names** into two new columns – Brand & Name
6. **Used One-Hot encoding** to convert categorical variables for use in modeling



MODELING PROCESS

BUILDING ACCURATE MODELS FOR OUR CUSTOMERS

ANALYZED CORRELATION BETWEEN INDEPENDENT AND TARGET VARIABLES…

Correlation between price and other variables:

- year: 0.280
- kilometers_driven: -0.011
- mileage: -0.304
- engine: 0.606
- power: 0.706
- new_price: 0.367

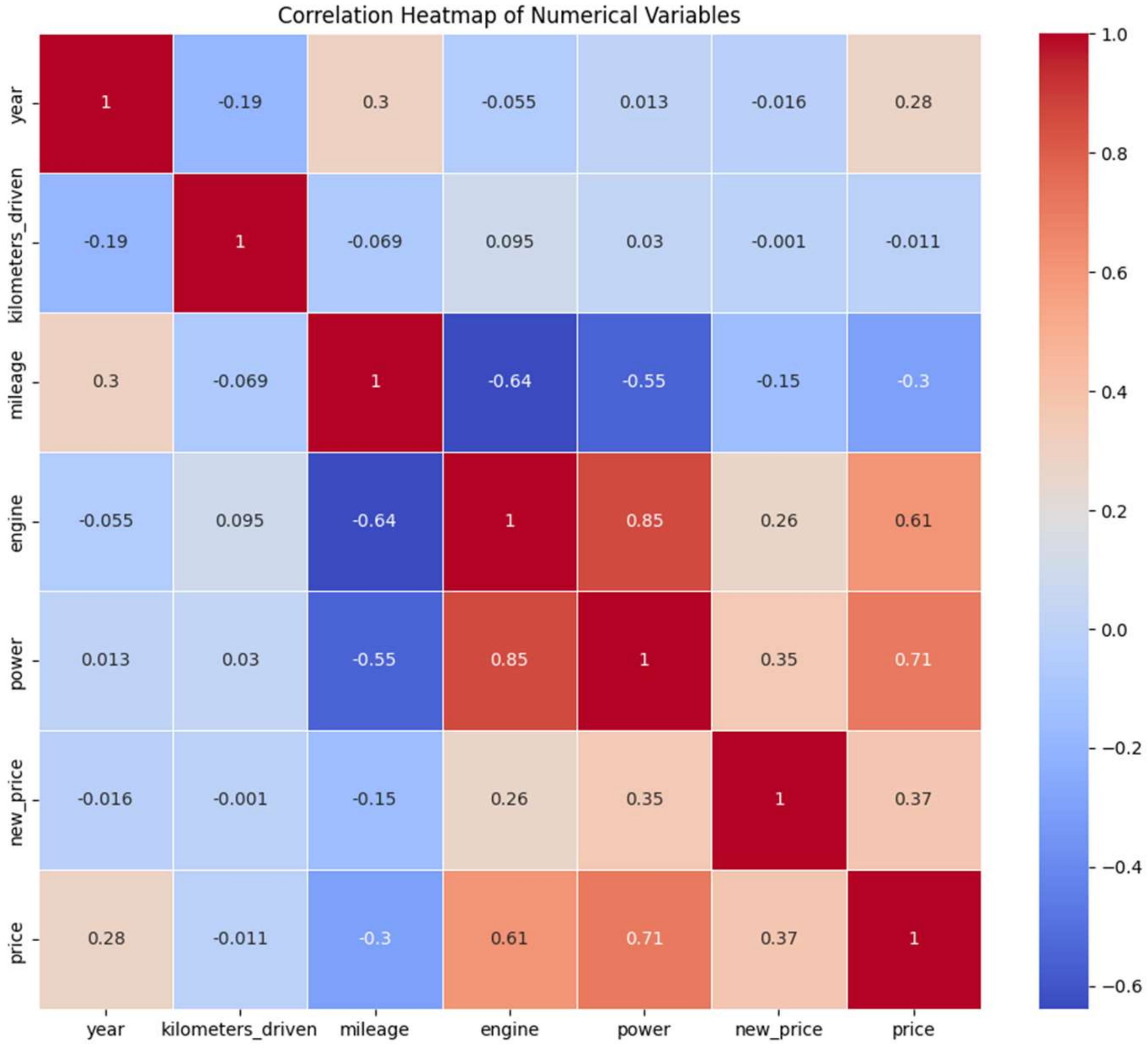
Potential multicollinearity issues:

High correlation (0.85) between 'engine' and 'power' - potential multicollinearity issue.

Other statistical observations:

- Highest positive correlation with price: 0.706 (power)
- Highest negative correlation with price: -0.011 (kilometers_driven)
- Variables with near-zero correlation with price: kilometers_driven

Correlation Heatmap of Numerical Variables



	year	kilometers_driven	mileage	engine	power	new_price	price
year	1	-0.19	0.3	-0.055	0.013	-0.016	0.28
kilometers_driven	-0.19	1	-0.069	0.095	0.03	-0.001	-0.011
mileage	0.3	-0.069	1	-0.64	-0.55	-0.15	-0.3
engine	-0.055	0.095	-0.64	1	0.85	0.26	0.61
power	0.013	0.03	-0.55	0.85	1	0.35	0.71
new_price	-0.016	-0.001	-0.15	0.26	0.35	1	0.37
price	0.28	-0.011	-0.3	0.61	0.71	0.37	1

STARTED WITH STANDARD LINEAR & OLS REGRESSION MODELS...

OLS Regression Results	
==	
Dep. Variable:	price_log R-squared: 0.750
Model: OLS	Adj. R-squared: 0.747
Method:	Least Squares F-statistic: 276.7
Date: Tue, 06 Aug 2024	Prob (F-statistic): 0.00
Time: 07:06:04	Log-Likelihood: -1773.9
No. Observations: 5040	AIC: 3658.
Df Residuals: 4985	BIC: 4017.
Df Model: 54	Covariance Type: nonrobust

PROCEEDED TO PERFORM ADDITIONAL REGRESSION MODELS

Metric	Linear Regression (standard)	OLS	Ridge	Decision Tree	Random Forest
R-squared Score	0.684112	0.748000	0.701062	0.550653	0.698109
Root Mean Squared Error	5.992148	0.377072	5.829170	7.146717	5.857886

In total, five (5) different modeling techniques were tested to find the most accurate approach for our customers…

Based on the table comparing the performance of the five models, the **OLS (Ordinary Least Squares) model appears to be performing relatively better than the other models.** It has the highest R-squared score of 0.748 and the lowest Root Mean Squared Error (RMSE) of 0.37707235524876515.

The **high R-squared score indicates that the OLS model explains a significant portion** of the variance in the target variable (used car prices), while the low RMSE suggests that the model's predictions have a relatively small average error compared to the actual values.

While the Random Forest model comes close in performance to the Ridge Regression model, neither surpasses the OLS model's performance in terms of both R-squared and RMSE.

However, **there is still scope for further improvement** in the model's performance. Even though the OLS model has the best performance among the five models, its R-squared score of 0.748 indicates that there is still some unexplained variance in the target variable that could potentially be captured by exploring additional features or using more advanced modeling techniques.

KEY FEATURES

Other Notes: Using the table of model coefficients and p-values that were derived using the best-performing OLS model, we can derive several meaningful insights about car pricing….

Positive Influences on Price:

- new_price: The coefficient is positive (1.335037e-03) with a very small p-value (2.427166e-02), indicating that higher original prices correlate with higher used car prices.
- engine: A positive coefficient (1.670184e-04) and a very small p-value (1.191925e-10) suggest that cars with more powerful engines tend to have higher prices.
- power: This also has a positive impact (2.955084e-05) with a very significant p-value (4.993497e-28), indicating that higher power ratings contribute to higher prices.
- location_Bangalore: A positive coefficient (6.901798e-02) with a relatively small p-value (3.149623e-02) suggests that cars in Bangalore may be priced higher than in other locations.

Negative Influences on Price:

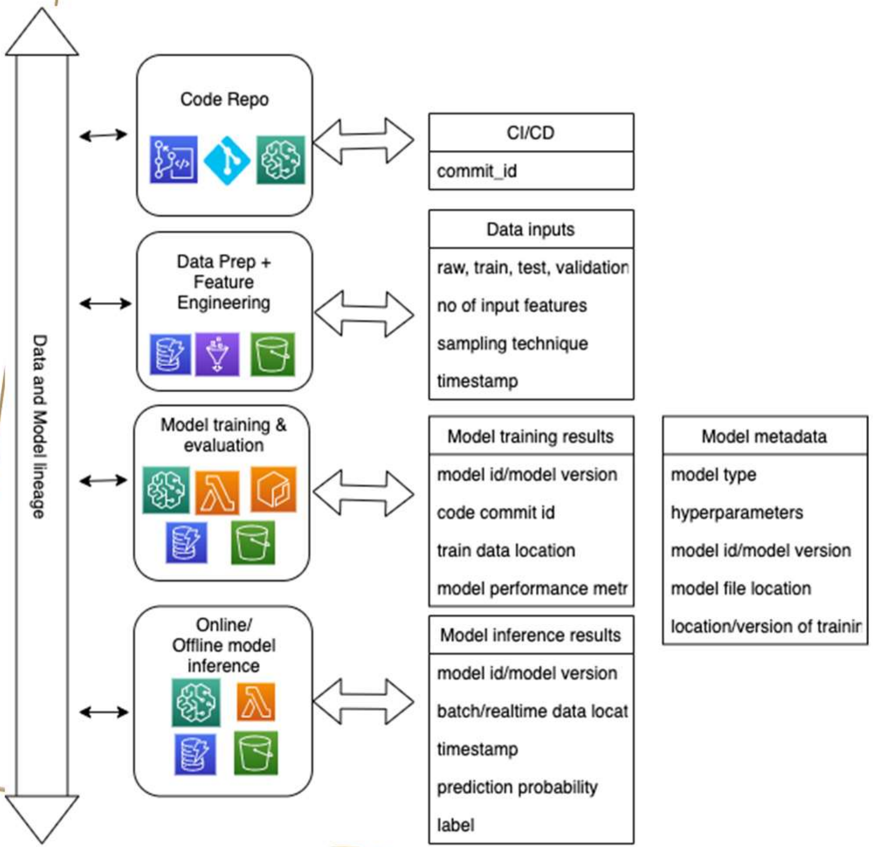
- mileage: A negative coefficient (-1.425920e-02) and a very significant p-value (9.349975e-10) indicate that higher mileage decreases the car price.
- kilometers_driven_log: A negative coefficient (-4.980287e-02) and a significant p-value (8.189808e-08) also suggest that more kilometers driven lowers the car price.
- transmission_Manual: This feature has a negative coefficient (-8.587928e-02) with a very small p-value (1.754207e-07), indicating that manual transmissions are associated with lower prices compared to automatic transmissions.
- owner_type_Second and owner_type_Third: Both have negative coefficients (-4.368074e-02 and -1.040836e-01 respectively) and significant p-values (2.869216e-03 and 5.868476e-03), suggesting that cars with multiple previous owners are priced lower.

We will use this data to present our pricing to our customers and instill TRUST in our platform…

NEXT STEPS

INTEGRATING OUR MODELS INTO OUR WEBSITE

KEY FEATURES


Data and Model lineage
Code Repo
Data Prep + Feature Engineering
Model training & evaluation
Online/ Offline model inference

CI/CD
commit_id

Data inputs
raw, train, test, validation
no of input features
sampling technique
timestamp

Model training results
model id/model version
code commit id
train data location
model performance metr

Model metadata
model type
hyperparameters
model id/model version
model file location
location/version of trainir

Model inference results
model id/model version
batch/realtime data locat
timestamp
prediction probability
label

After successfully developing the initial draft of our machine learning model for predicting used car prices, we are poised to take a significant leap forward in our project. The next phase involves **productization of the model**, a critical step where we transition from a prototype to a fully operational service. To achieve this, we will be **integrating our model with AWS SageMaker**, a robust and scalable platform that will not only streamline deployment but also facilitate the management of our machine learning lifecycle.

By leveraging SageMaker's capabilities, we **will expose a predictions API that will seamlessly integrate with our website**. This will allow our customers to obtain real-time, accurate price predictions for used cars, enhancing their user experience and our platform's value proposition.

Furthermore, we recognize that our current model, as it stands, **is a starting point.** Machine learning models thrive on continuous learning and improvement. Therefore, **we will utilize SageMaker's powerful experimentation and tracking features to conduct further training and enhancements to improve our model over time**.

As we progress, we will maintain a keen focus on tracking the model's evolution over time. SageMaker's robust versioning and tracking systems will enable us to monitor improvements and regressions, ensuring that only the best-performing models are deployed. We plan to release new and improved versions of our model to our website periodically, aligning with our commitment to deliver excellence and drive innovation in the used car market.

RISKS AND CHALLENGES

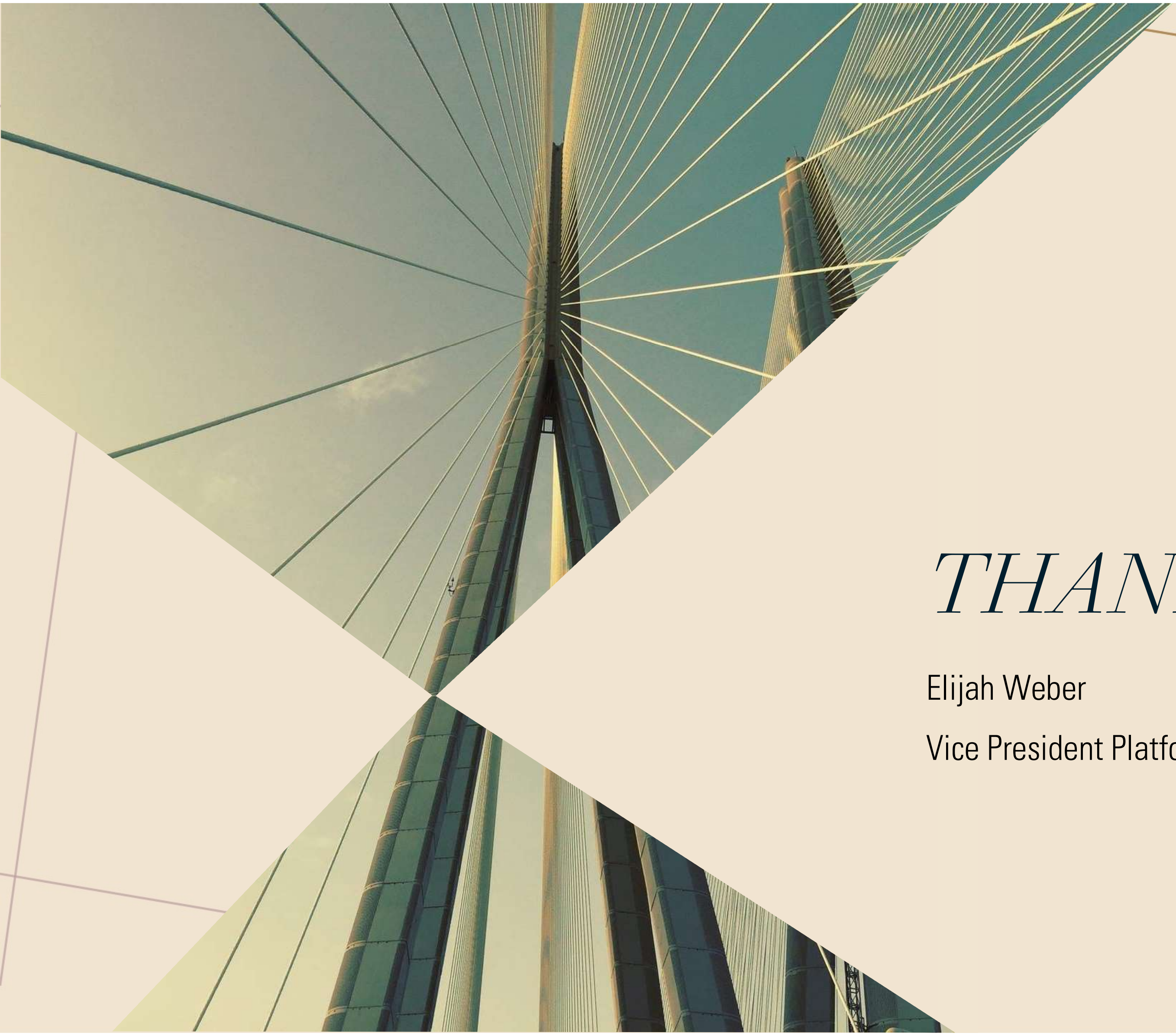
Risks and Challenges:

- **Market Competition:** The online used car market is a “Red Ocean” environment, with established players already having a strong foothold.
- **Data Security:** Protecting sensitive customer data and transaction details is paramount.
- **Technology Scalability:** Ensuring the platform can handle increased traffic and data volume as the business grows.
- **User Trust:** Building trust with users in terms of fair pricing, car condition, and transaction security.
- **Model Accuracy:** Developing a machine learning model that accurately predicts car prices and provides value to users.

Red Ocean Strategy vs Blue Ocean Strategy | Learn the Difference

Addressing Challenges with Scientific Approaches:

- **Data-Driven Design:** Utilize A/B testing and user feedback to iteratively improve the website’s user interface and user experience.
- **Scalable Architecture:** Design the technology stack to be scalable using cloud services and microservices architecture.
- **Transparency:** Foster user trust by being transparent about pricing algorithms and providing detailed car histories.
- **Continuous Model Improvement:** Apply scientific methods such as hypothesis testing and experimental design to refine the machine learning model. Use techniques like cross-validation and feature engineering to enhance model performance.



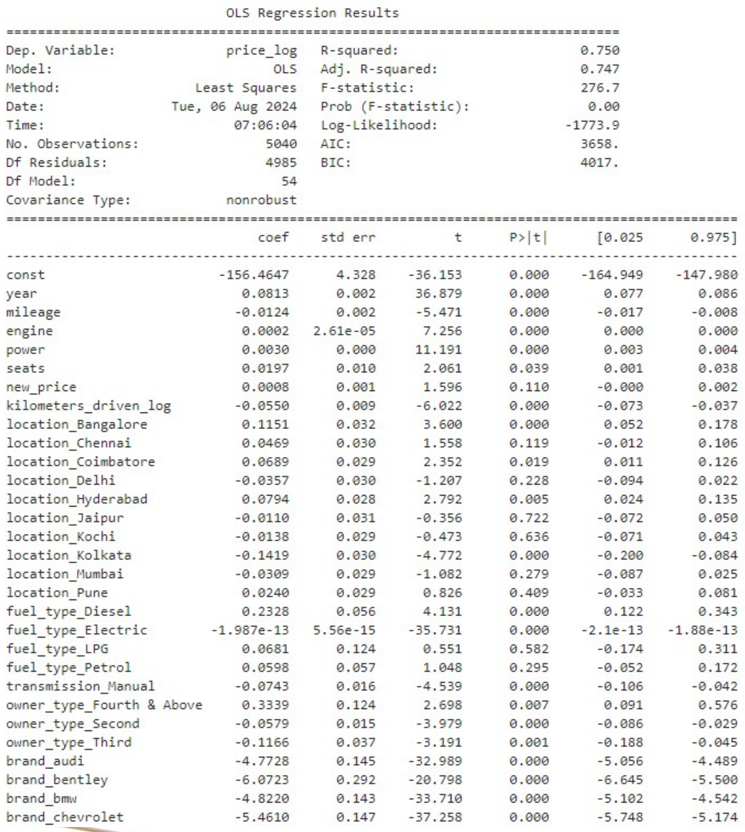
THANK YOU

Elijah Weber

Vice President Platform Solutions Architecture

APPENDIX 1

OLS MODEL RESULTS



```
                            OLS Regression Results
==============================================================================
Dep. Variable:             price_log   R-squared:                       0.750
Model:                           OLS   Adj. R-squared:                  0.747
Method:                Least Squares   F-statistic:                     276.7
Date:               Tue, 06 Aug 2024   Prob (F-statistic):               0.00
Time:                       07:06:04   Log-Likelihood:                -1773.9
No. Observations:               5040   AIC:                             3658.
Df Residuals:                   4985   BIC:                             4017.
Df Model:                         54
Covariance Type:           nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                       -156.4647      4.328    -36.153      0.000    -164.949    -147.980
year                           0.0813      0.002     36.879      0.000       0.077       0.086
mileage                       -0.0124      0.002     -5.471      0.000      -0.017      -0.008
engine                         0.0002   2.61e-05      7.256      0.000       0.000       0.000
power                          0.0030      0.000     11.191      0.000       0.003       0.004
seats                          0.0197      0.010      2.061      0.039       0.001       0.038
new_price                      0.0008      0.001      1.596      0.110      -0.000       0.002
kilometers_driven_log         -0.0550      0.009     -6.022      0.000      -0.073      -0.037
location_Bangalore             0.1151      0.032      3.600      0.000       0.052       0.178
location_Chennai               0.0469      0.030      1.558      0.119      -0.012       0.106
location_Coimbatore            0.0689      0.029      2.352      0.019       0.011       0.126
location_Delhi                -0.0357      0.030     -1.207      0.228      -0.094       0.022
location_Hyderabad             0.0794      0.028      2.792      0.005       0.024       0.135
location_Jaipur               -0.0110      0.031     -0.356      0.722      -0.072       0.050
location_Kochi                -0.0138      0.029     -0.473      0.636      -0.071       0.043
location_Kolkata              -0.1419      0.030     -4.772      0.000      -0.200      -0.084
location_Mumbai               -0.0309      0.029     -1.082      0.279      -0.087       0.025
location_Pune                  0.0240      0.029      0.826      0.409      -0.033       0.081
fuel_type_Diesel               0.2328      0.056      4.131      0.000       0.122       0.343
fuel_type_Electric         -1.987e-13   5.56e-15    -35.731      0.000    -2.1e-13   -1.88e-13
fuel_type_LPG                  0.0681      0.124      0.551      0.582      -0.174       0.311
fuel_type_Petrol               0.0598      0.057      1.048      0.295      -0.052       0.172
transmission_Manual           -0.0743      0.016     -4.539      0.000      -0.106      -0.042
owner_type_Fourth & Above      0.3339      0.124      2.698      0.007       0.091       0.576
owner_type_Second             -0.0579      0.015     -3.979      0.000      -0.086      -0.029
owner_type_Third              -0.1166      0.037     -3.191      0.001      -0.188      -0.045
brand_audi                    -4.7728      0.145    -32.989      0.000      -5.056      -4.489
brand_bentley                 -6.0723      0.292    -20.798      0.000      -6.645      -5.500
brand_bmw                     -4.8220      0.143    -33.710      0.000      -5.102      -4.542
brand_chevrolet               -5.4610      0.147    -37.258      0.000      -5.748      -5.174
```

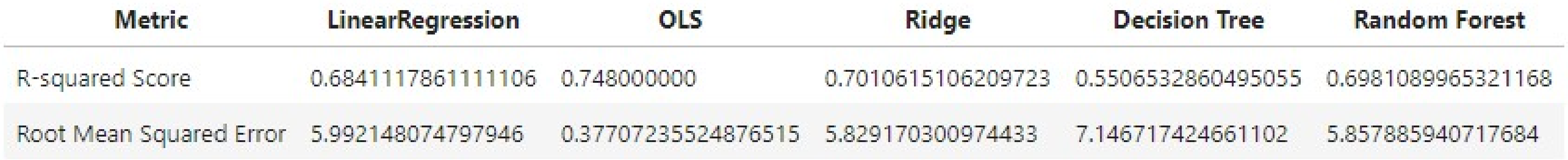
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The smallest eigenvalue is 3.37e-22. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Most overall significant categorical varaibles of LINEAR REGRESSION are :
['seats', 'location', 'owner_type', 'fuel_type', 'transmission', 'mileage', 'kilometers_driven', 'kilometers_driven_log', 'engine', 'brand', 'power', 'year']

MODEL COMPARISON

- PERFORMANCE BETWEEN THE Five MODELS



Metric	LinearRegression	OLS	Ridge	Decision Tree	Random Forest
R-squared Score	0.6841117861111106	0.748000000	0.7010615106209723	0.5506532860495055	0.6981089965321168
Root Mean Squared Error	5.992148074797946	0.37707235524876515	5.829170300974433	7.146717424661102	5.857885940717684

- Of the Five, the OLS Model **still** appears to have the best performance with the highest R2 & lowest RMSE.
- While the Random Forest Model is MUCH closer in performance to the ridge regression, it still does not surpass the OLS performance.