Demystifying Machine Learning
Craigslist Used Cars Prices

Group 2

- Fahmida Billa
- Gustavo Mendes
- Kevin Mosweu
- Shankar Mohanathas

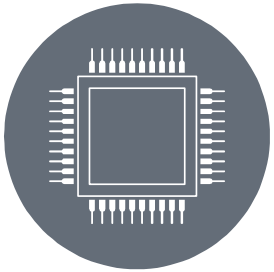
The Problem

To develop a machine learning model that can accurately analyze the prices of used vehicles listed on Craigslist, creating a predictive model that will assist both sellers and buyers in making informed decisions regarding vehicle pricing.

Expected Outcomes

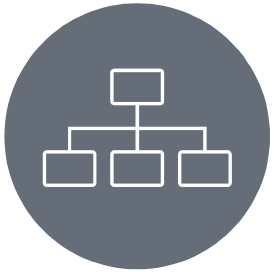
Insights

Insights into the factors that significantly affect vehicle prices and their relative importance



ML

A machine learning model capable of accurately predicting the prices of used vehicles listed on Craigslist



Dashboard

An interactive application that enables users to see trends in the data for the top five important features

PREPARING THE DATA

Data Preprocessing:

- Initial dataset: 26 columns, 440,000+ rows
- Read into a Pandas DF using Spark
- Conducted thorough preprocessing:
    - Removed duplicates
    - Handled missing and null values
    - Addressed inconsistencies in the dataset
    - Engineer additional features, such as age of the vehicle, which can enhance the predictive power of the model.

Final Dataset (head):

	price	manufacturer	model	condition	cylinders	fuel	odometer	transmission	drive	type	state	car_age
0	33590	gmc	others	good	8 cylinders	gas	57923.0	other	unknown	pickup	al	7
1	22590	chevrolet	silverado	good	8 cylinders	gas	71229.0	other	unknown	pickup	al	11
2	39590	chevrolet	others	good	8 cylinders	gas	19160.0	other	unknown	pickup	al	1
3	30990	toyota	others	good	8 cylinders	gas	41124.0	other	unknown	pickup	al	4
4	15000	ford	others	excellent	6 cylinders	gas	128000.0	automatic	rwd	truck	al	8

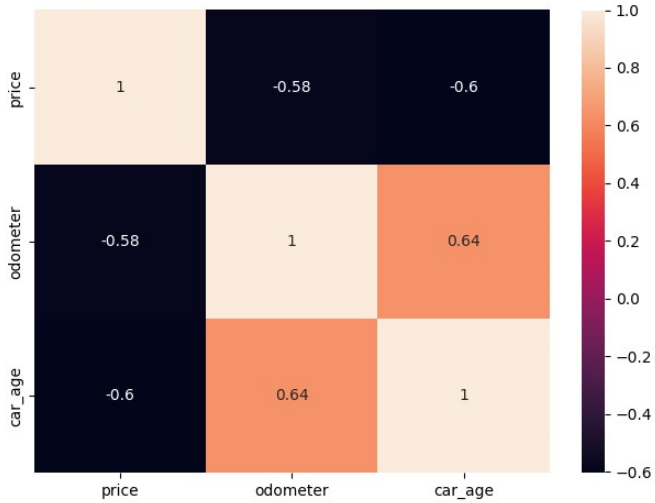
258627 rows

MODELING THE DATA

Data Modeling:

•Perform exploratory data analysis to gain insights into the distribution of features, identify outliers, and understand potential correlations.

•Identify relevant features from the dataset, such as price, condition, manufacturer, state, and other categories that influence vehicle prices.



Numeric Features Correlation - Heatmap

TRAINING THE DATA

Scaling and Training Data into different ML models

Compare the performance of the model with baseline models and industry standards to gauge its effectiveness.

- Linear Regression Model

```
R2_score:  0.7403
Mean squared error:  34549047.32
Mean absolute error:  4398.06
Root mean squared error:  5877.84
```

Lasso Regression

```
R2_score:  0.7404
Mean squared error:  34544079.71
Mean absolute error:  4397.41
Root mean squared error:  5877.42
```

•Ridge Regression

```
R2_score:  0.7404
Mean squared error:  34544089.32
Mean absolute error:  4397.41
Root mean squared error:  5877.42
```

- Random Forest Regression

```
R2 score:  0.9288
Mean squared error:  9475684.23
Mean absolute error:  1611.74
Root mean squared error:  3078.26
```

- Decision Tree Model

```
R2 score:  0.8676
Mean squared error:  17614194.52
Mean absolute error:  1893.04
Root mean squared error:  4196.93
```

- Elastic Net Regression

```
R2_score:  0.7403
Mean squared error:  34546477.28
Mean absolute error:  4396.77
Root mean squared error:  5877.63
```

TRAINING THE DATA

Scaling and Training Data into different ML models

Evaluate the trained model using relevant evaluation metrics, such as mean absolute error (MAE), mean squared error (MSE), and R-squared.

	Linear Regression	Lasso Regression	Ridge Regression	Elastic Net Regression	Random Forest Regression	Decision Tree Regression
R2 Score	0.7403	0.7404	0.7404	0.7403	0.9288	0.8676
Accuracy(%)	74.0320	74.0357	74.0357	74.0339	92.8778	86.7607
Mean Squared Error	34549047.32	34544079.71	34544089.32	34546477.28	9475684.23	17614194.52
Mean Absolute Error	4398.06	4397.41	4397.41	4396.77	1611.74	1893.04
Root MSE	5877.84	5877.42	5877.42	5877.63	3078.26	4196.93

Model Evaluation:

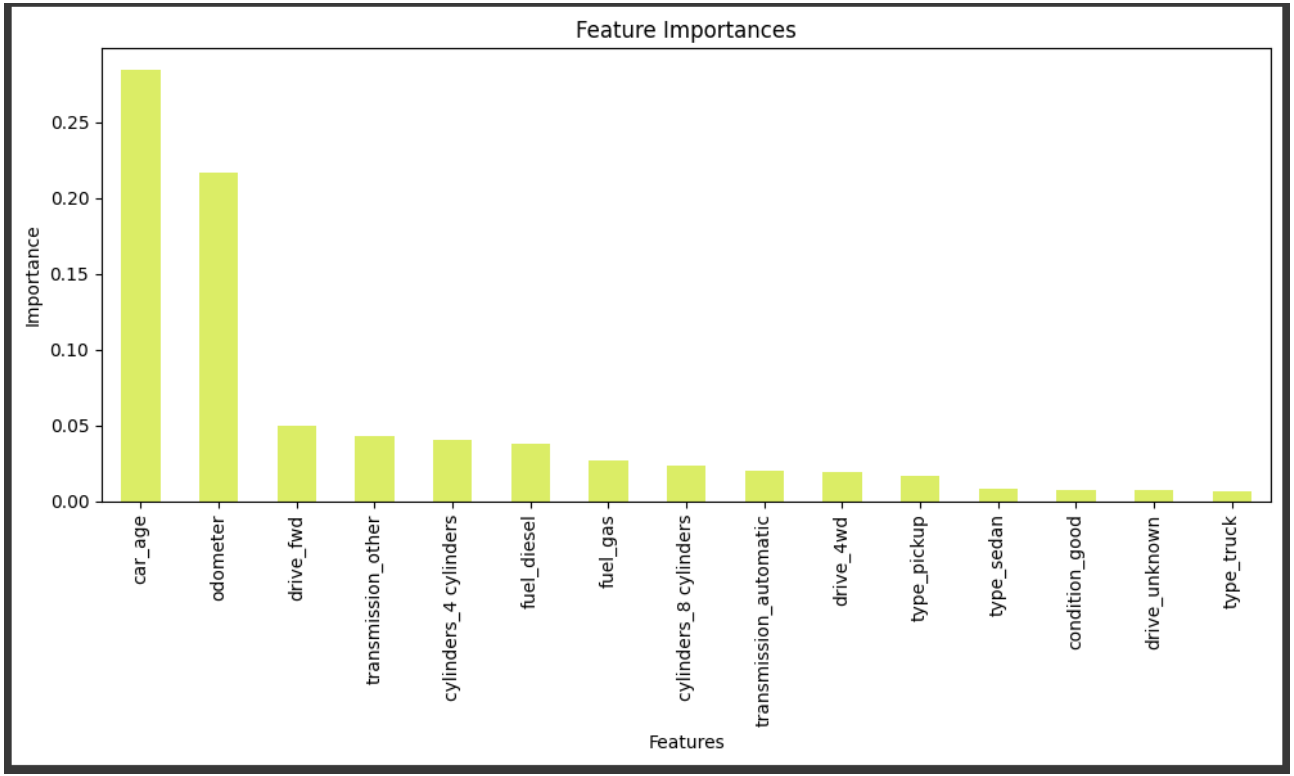
- Random Forest Regression model outperformed other models, achieving an R2 score of 0.92, indicating strong predictive power.
- Attained an accuracy of 92.88%, demonstrating high precision in price predictions.

Best Parameters for Random Forest

The best peforming Random Tree Regressor parameters:
n_estimators=300
random_state=0
min_samples_leaf=1
max_features=0.3
n_jobs=-1
oob_score=True

```
R2 score:  0.9290
Mean squared error:  9442190.55
Mean absolute error:  1607.26
Root mean squared error:  3072.81
```

The Most Important Features



Statistical Tests

Linear Regression : Car Age and Odometer

Linear regression done for price against the car age revealed an r value of -0.60 and p value of 0.0.

Linear regression done for price against the odometer revealed an r value of -0.58 and p value of 0.0.

T-Tests Evaluation: The T-Test P-Values of Drive, Cylinders and Fuel in regard to average price were less than 0.05 .

TABLEAU

Use this QR code to access our Tableau Dashboard

CONCLUSION

- Random Forest machine learning model was found to produce the best r-squared value of 0.929

- With this being said, our model provides a good prediction of vehicle listing price on Craigslist

QUESTIONS?