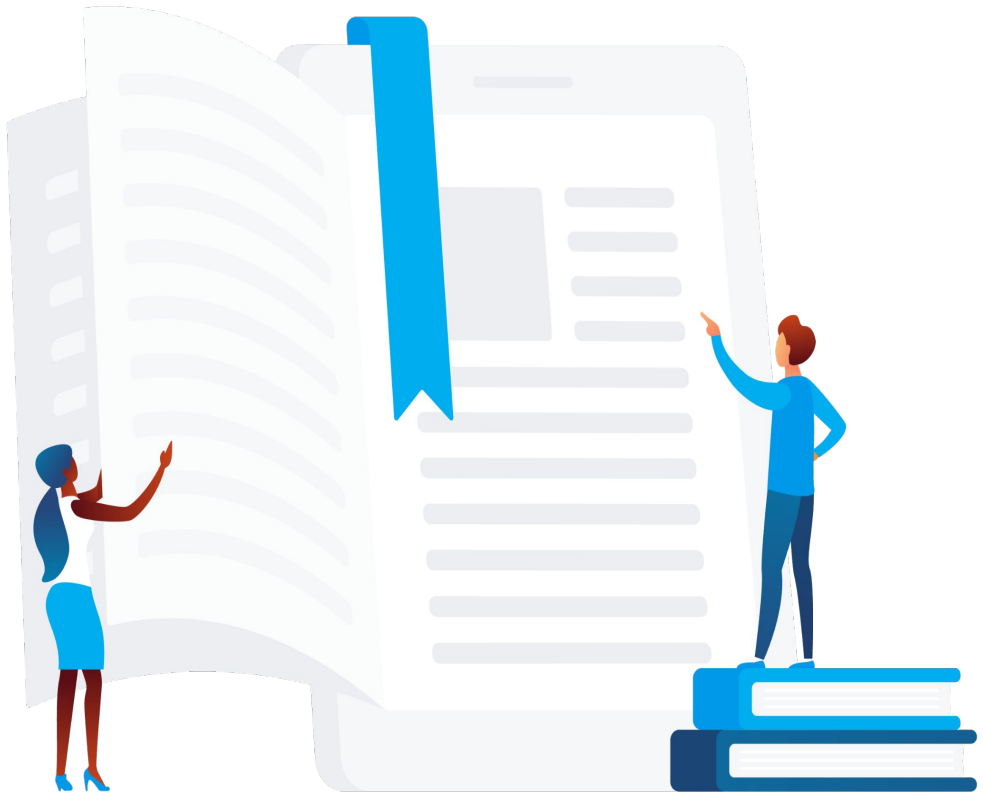
Used Car Price Predictions

Team 15 - MSCI 546 Project

Meenakshi Andoorveedu (20837987)
Nayeema Nonta (20837920)
Peter Twarecki (20849709)
Joanna Yang (20826548)

Topic and Value Proposition

Predicting the price of used cars using the US Used Cars Dataset [1]



Predictions for Prospective Buyers

The model serves as a tool for buyers interested in understanding the key factors that influence car prices and assists them in making purchasing decisions

Relevance to Current Trends

The used car market had supply shortages during the pandemic and is projected to recover to normal in 2024 [2]. The model provides insights for researchers who wish to analyze this market and the trends that occured

Financial Services and Insurance

Financial institutions and insurance companies can use price predictions to assess vehicle value for loan and insurance policies, ensuring more accurate risk assessment and pricing

Machine Learning Task

Type of ML Task Implemented

Machine learning task is regression since we are predicting the price of used cars, which is a continuous variable

Regression utilizes features such as make, model, mileage, etc. to predict the car's price

Classification tasks predict discrete labels, which is not the case for this task

Data Overview

Data Description, Feature Engineering, and Cleaning

Original Dataset

Columns: 66, including 32 string, 8 decimal, and others

Rows: 3,000,599

Sample Columns: body_type, engine_cylinders, exterior_color, fleet, highway_fuel_economy, mileage, transmission, wheelbase, year

Feature Engineering

One-Hot Encoding: Categorical variables like body_type were one-hot encoded, since there were only 9 distinct categories (sedan, coupe, SUV, etc.)

Max-Min Normalization: For the neural network model, all the data were scaled to values between 0-1, to ensure weightings were more equally distributed

Data Cleaning

1. 19 columns removed to reduce size of dataset, including identification columns like vin. Some columns were similar to others, i.e. transmission very similar to transmission_display, which was removed

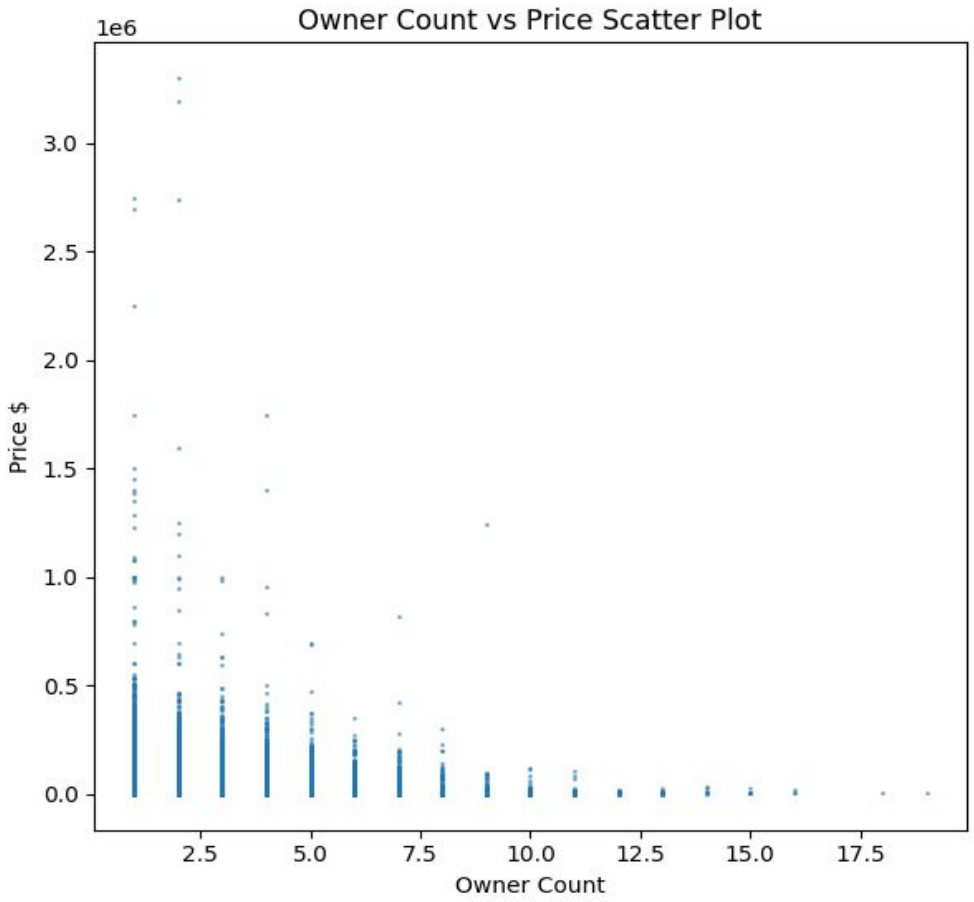
2. 11 additional columns were removed, the majority of which were binary and had 48% of their records being null, bringing the total to 36 columns

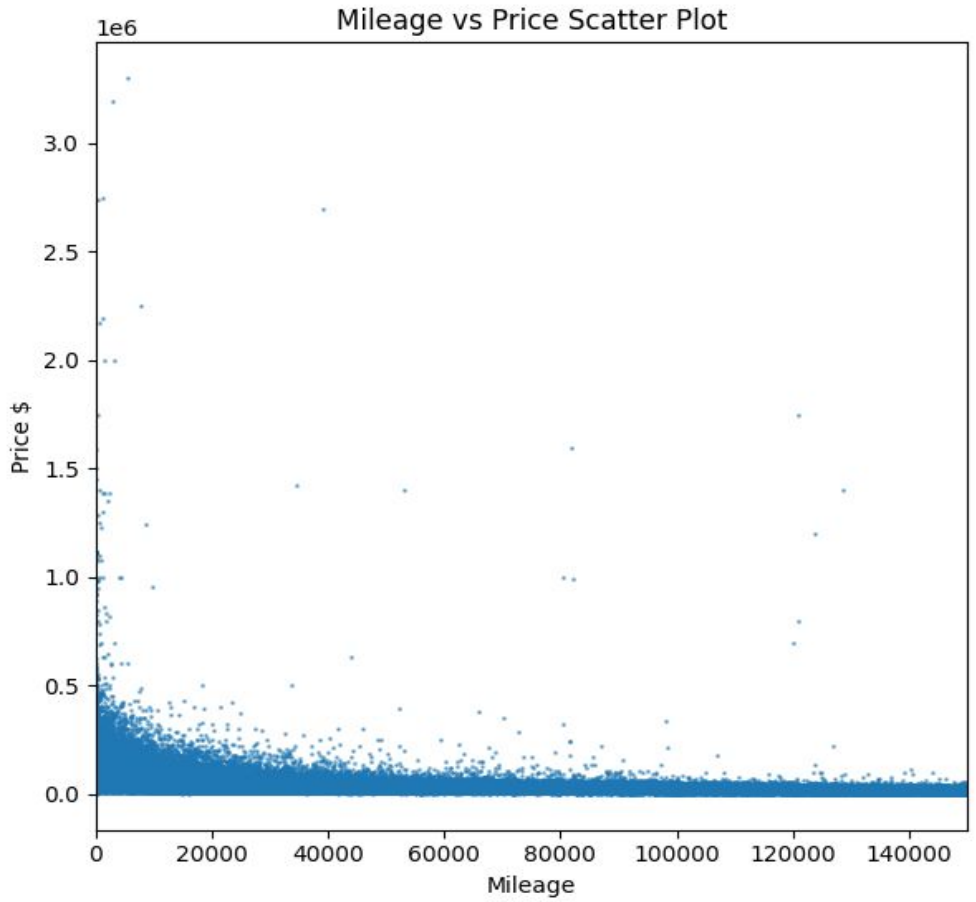
3. Columns like the height and weight of the car were formatted as decimals, with the word " in" to represent inches after them. Any unnecessary strings were removed from continuous variables

4. Categorical variables with too many columns were also removed (tens, hundreds, or more), these included city, dealer_zip, interior_color, listed_date, major_options, and model_name

Exploratory Data Analysis (Numeric)

Breakdown of Numeric Data





Owner Count vs. Price

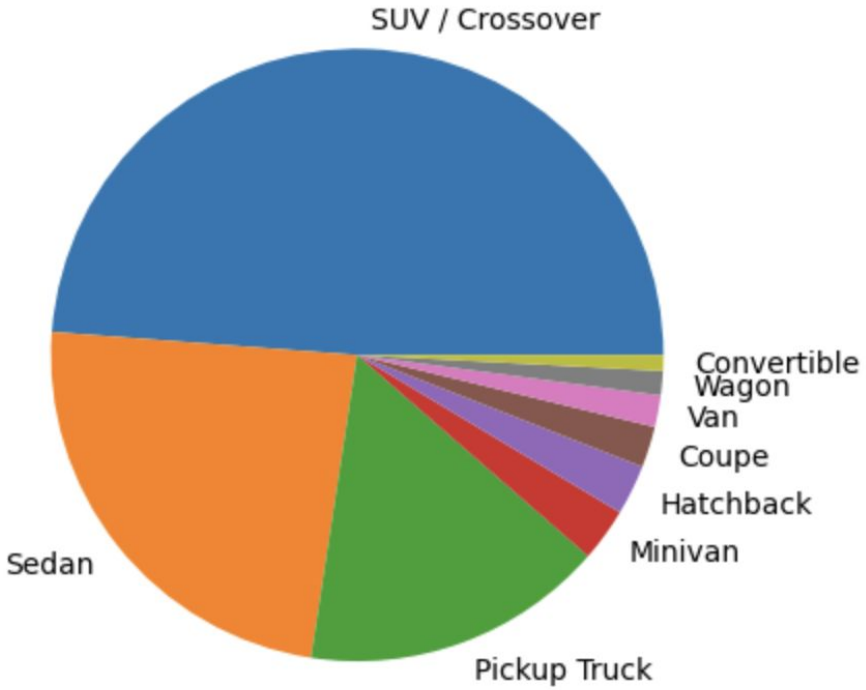
Illustrates an exponential decrease in price with rising owner count, which respects logical reasoning since if a car has more owners, the price of the car would likely depreciate

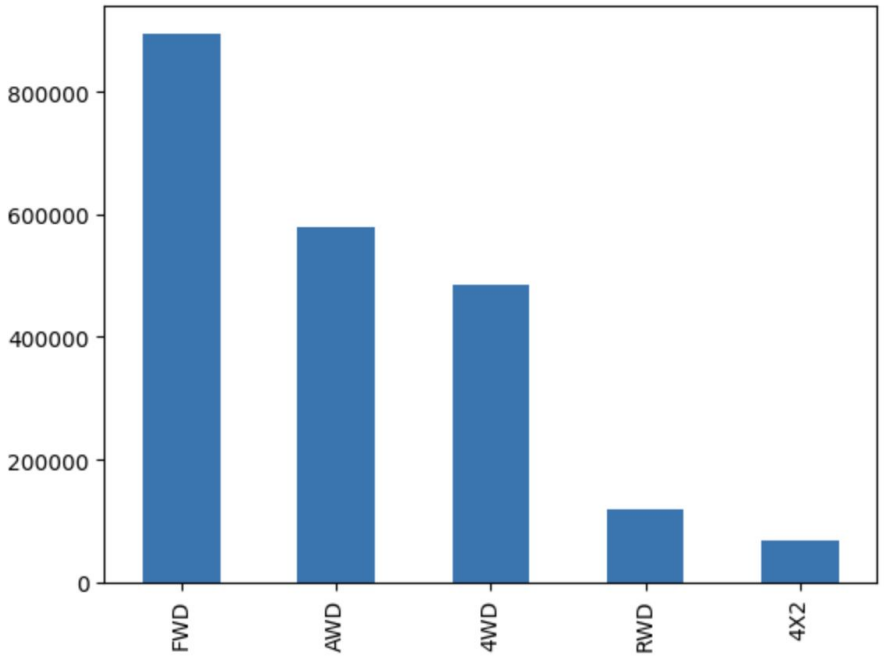
Mileage vs. Price

The scatter plot exhibits a slight exponential decrease in price as mileage increases. This suggests the distance that a car has been driven has a slight negative correlation with price

Exploratory Data Analysis (Categorical)

Breakdown of Categorical Data





Body Type

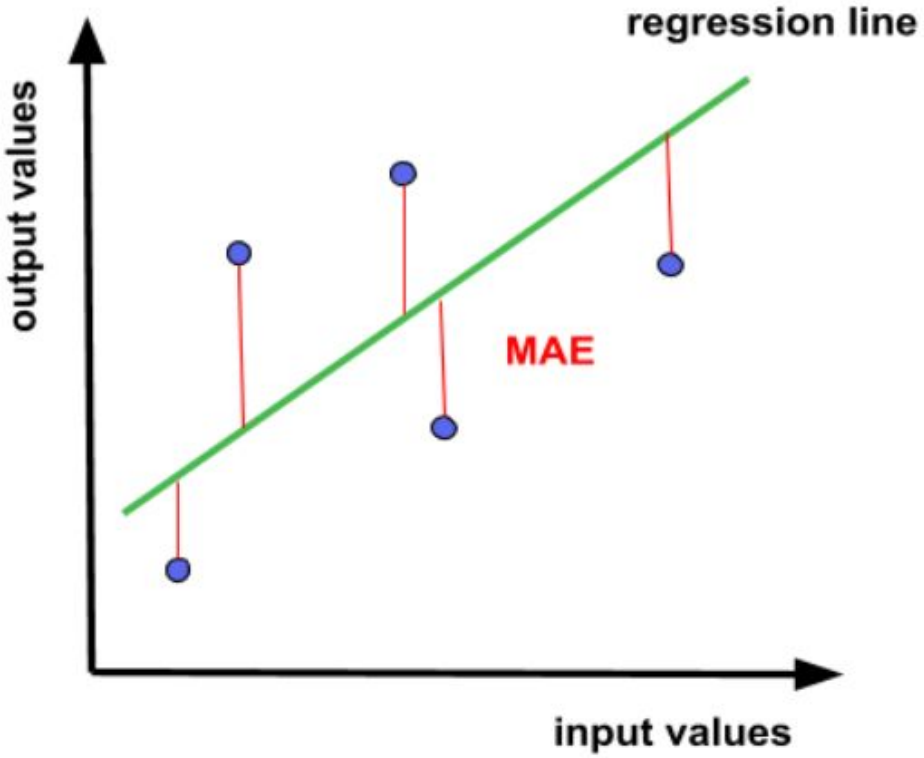
The plurality of the dataset are SUV/Crossover vehicles, with over 80% being either SUVs, Sedans, and Pickup Trucks

Wheel System (or Drivetrain)

Most of the vehicles in the dataset are FWD, which are offered on most discount or inexpensive vehicles, while AWD or RWD are typically offered in more luxurious or sports cars

Performance Metrics

Metrics and Rationale



Coefficient of Determination – R^2

Measures the **goodness of fit** which ranges from 0 to 1 describing the proportion of variance of the dependent variables that is explainable by the independent variables

Mean Squared Error (MSE)

Determines the squared difference between the predicted and true values. It should be noted that this metric **penalizes larger errors** and is sensitive to outliers [3]

Root Mean Squared Error (RMSE)

Measures the euclidean distance between predicted and true values. RMSE is **sensitive to outliers** but is highly interpretable due to it being in the same units as the data

Mean Absolute Error (MAE)

Measures the difference in magnitude of the predicted and actual values. MAE is **robust to outliers** and is also highly interpretable due to it being in the same units as the data [3]

Linear Regression (baseline)

Insights & Rationale

Why Linear Regression?

STRENGTHS

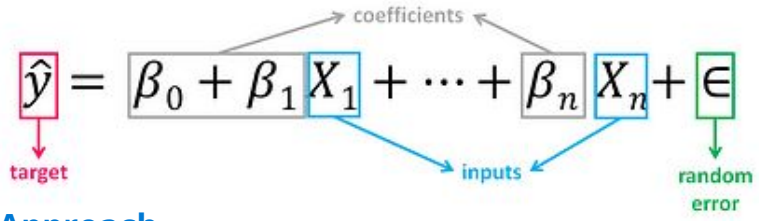
- Easy to understand and **interpret results** [4]
- **Less computationally expensive** and easy to train [5]

WEAKNESSES

- Tends to oversimplify real-world problems by **assuming a linear relationship** between features and target variables [6]

How it Works

Linear regression mathematically models a linear equation between independent and dependent variables to minimize the error [7]



Approach

FEATURE SETS & PARAMETERS

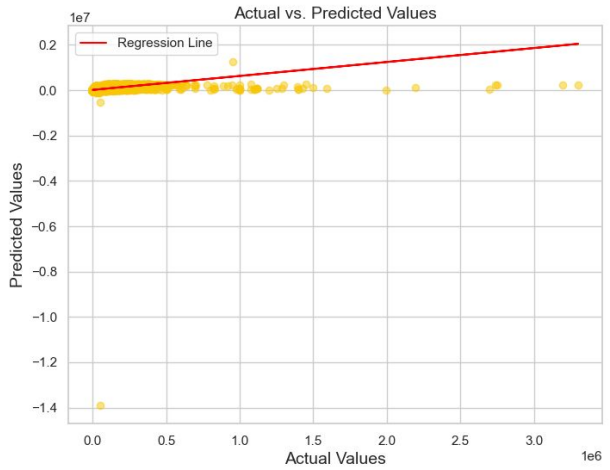
- 40 features were used to train the baseline model
- No hyperparameters were used to ensure that the model represents a simple base case, providing a benchmark for comparison

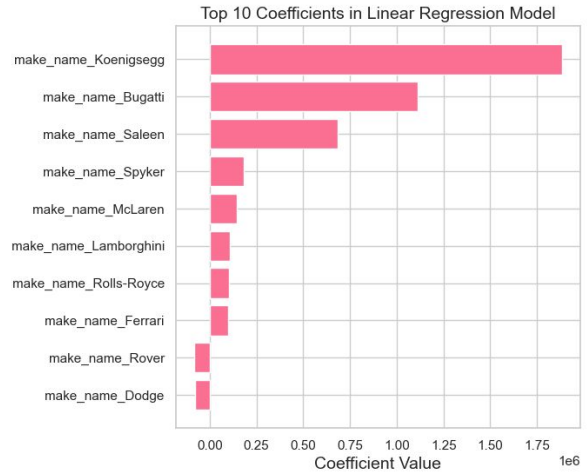
Insights

RESULTS & INSIGHTS FROM BEST ITERATION

The feature set used in iteration 1 (base model features) had the best results. **Baseline** metrics will be compared to four machine learning models.

R^2	MSE	MAE	RMSE
0.338154	$256,894,726.68	$6,029.57	$16,027.93





1) The model has **very low accuracy** specifically for the outliers (expensive cars) as it appears that the model predicts a similar price for cars that vary greatly in price

2) The make of the car has the **highest positive coefficient** value (ex. *make_name_Bugatti*) signifying that as the value of these independent variables increases, the price of the car increases as well

Ridge Regression

Insights & Rationale

Why Ridge Regression?

STRENGTHS

- Handles **high correlations** between features by introducing penalty term [11]
- Prevents model from **overfitting** [11]

WEAKNESSES

- Model **loses interpretability** [12]
- Selection of the **hyperparameter** alpha is difficult [12]

How it Works

Ridge regression is a technique that reduces overfitting in machine learning models by addressing high correlations and penalizing large parameter weights, thus improving model accuracy.

Approach

FEATURE SETS & PARAMETERS

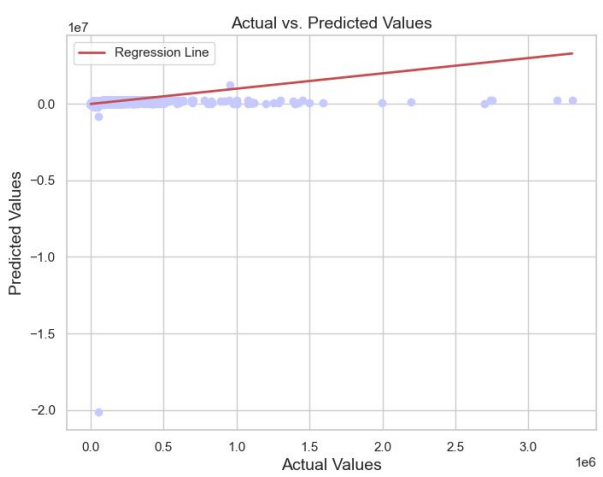
- Hyperparameters tested**:** alpha = 1, alpha = 10, alpha = 100, alpha = 0.01
  - Found alpha = 0.01 gave best results
- 4 iterations of feature subsets were performed where alpha = 0.01

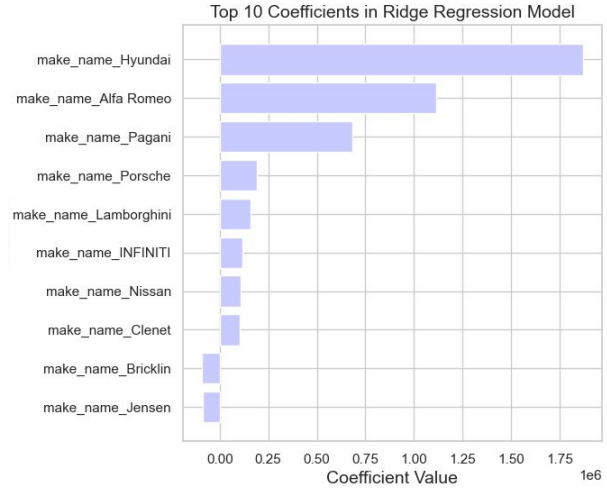
Insights

RESULTS & INSIGHTS FROM BEST ITERATION

The feature set used in (base model features with alpha = 0.01 had the best results (**slight improvement** in every metric from linear regression baseline).

R^2	MSE	MAE	RMSE
0.3381585	$256,892,789.16	$6,029.57	$16,027.88





1) Most data points are concentrated near the lower end of the range, and shows as the actual values increase, the **predicted values get worse**

2) **Hyundai** has the highest coefficient value, indicating it has the most significant linear relationship with price in this model

3) Results are very **similar to linear regression**

Random Forest

Insights & Rationale

Why Random Forest?

STRENGTHS

- Learns **complex relationships** and non-linear decision boundaries
- Performs well on large datasets with high dimensionality, meaning it **scales well** for an industry context like predicting used car prices [8]

WEAKNESSES

- Provides **less interpretability** compared to a decision tree and tends to be **computationally expensive** (i.e. requires a lot of memory) [9]

How it Works

Random forests are a meta-estimator and uses an ensemble learning method for regression [10]

It aggregates the results of multiple decision trees trained on random subsets of data, where the model outperforms any single decision tree [10]

Approach

FEATURE SETS & PARAMETERS

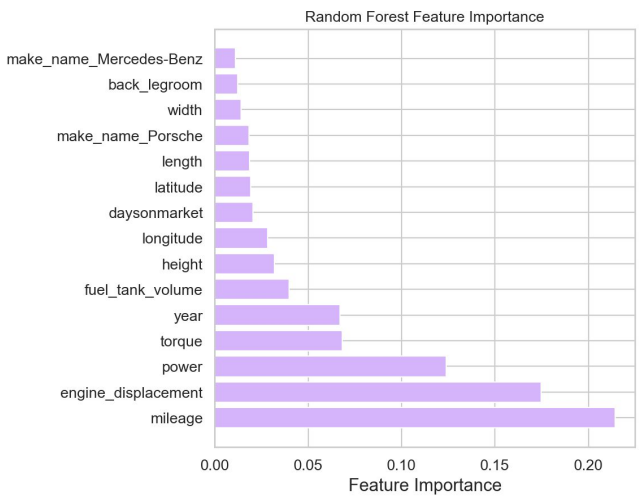
- 4 iterations of feature subsets were performed
- Hyperparameters were selected based on run-time feasibility & accuracy
    1. Base model: n_estimators = 10 (the number of decision trees)
    2. Random hyperparameter grid: max_depth = [10 to 100] max_leaf_nodes = [10 to 100]

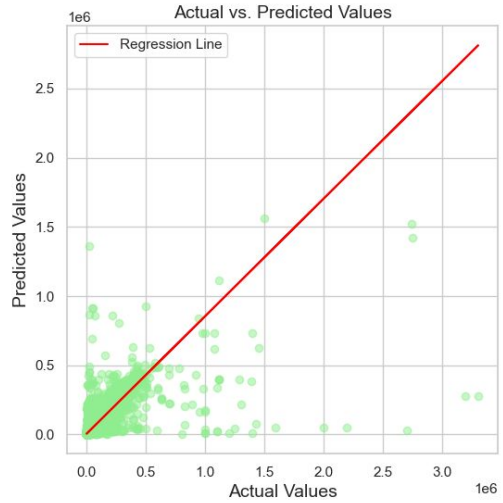
Insights

RESULTS & INSIGHTS FROM BEST ITERATION

The feature set used in iteration 1 (base model features) had the best results. **Improvement** in every metric from linear regression baseline.

R^2	MSE	MAE	RMSE
0.8436092	$60,702,856.60	$2,290.29	$7,791.20





1) **mileage** has the highest feature importance, followed by **engine displacement**, signifying that how far the car can be driven has the highest influence on a car's price

2) Since there is a **tight cluster of points** around the regression line, it can be observed that the model is relatively predicting the used cars' sale price accurately

3) However, since the points are concentrated at the bottom left corner, it means that the model **performs poorly for predicting higher prices**

XGBoost

Insights & Rationale

Why XGBoost?

STRENGTHS

- Provides insights into **feature importance**
  - Weight/frequency & gain (improvement in accuracy by feature) [13]
- Performs well on **structured and large datasets**

WEAKNESSES

- **Time consuming** for large complex data, requires more memory due to tree structure [14][15]
- Can disproportionately **focus on outliers** [13], which makes it bad for our dataset

How it Works

XGBoost for regression finds the optimal splits to minimize error during the construction of decision trees [16]

It evaluates all possible split points for **each feature** and selects the one that maximizes the reduction in MSE, referred to as the "**gain**" [17]

Approach

FEATURE SETS & PARAMETERS

- 4 iterations of feature subsets were performed
- Hyperparameters were selected based on **run-time** feasibility & **accuracy**
  - colsample_bytree: 0.8, boosting_iterations: 50
  - learning_rate: 0.01, max_depth: 5, alpha (L1 regularization): 10

Insights

RESULTS & INSIGHTS FROM BEST ITERATION

The feature set used in iteration 1 (base model features) had the best results. **Improvement** in every metric from linear regression baseline.

R^2	MSE	MAE	RMSE
0.3933524262	$235,901,549.89	$9,061.31	$15,359.09

1) **mileage** was a top feature to predict the price in terms of both weight and gain

2) Although **owner_count** is a less frequent feature, it is more important relative to other features as it is 3rd most important in terms of gain

3) daysonmarket, sp_id, savings_amount, and longitude all show lower importance but removing them resulted in a **1% increase in MSE**

Feed-forward Neural Network

Insights & Rationale

Why Feedforward Neural Network?

STRENGTHS

- Performs well on structured and large datasets
- Can handle **non-linear** and more complex relationships that a linear regression model can't [18]

WEAKNESSES

- **Time consuming** compared to classical machine learning models like linear regression or random forest, due to the higher number of weights that need to be learned [18]
- Black Box: why the neural networks perform well is **hard to interpret** [18]

Architecture & Parameters

- Input tensor of length 149, for the 149 columns in the dataset (includes one-hot encoded columns)
- 3 hidden layers, of size 200, 100, and 25, respectively.
- **Activation Function between Layers**: rectified linear unit (ReLU)
- **Learning rate**: 0.001
- **Optimization algorithm**: Adam
- **Loss function**: MSE
- **Training Dataset Size**: 0.4, or just over 1.2 million data records

Insights

RESULTS & INSIGHTS FROM BEST ITERATION

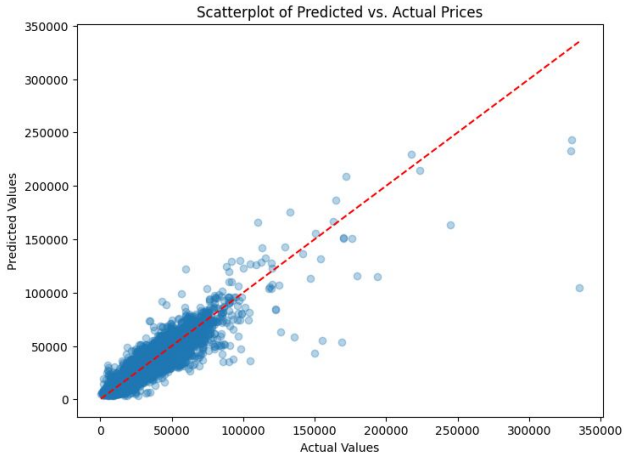
The neural network was iterated over 10 epochs. Each epoch took around 2.5 minutes to train.

The results of the final epoch are listed in the table below:

R^2	MSE	MAE	RMSE
0.78642405	$80,625,872.00	$3,788.42	$8,979.19

This network was trained on max-min normalized data, or data transformed between the values of 0 and 1. Having tested the same network architecture on un-normalized data, the results are slightly worse:

R^2	MSE	MAE	RMSE
0.735763702	$99,750,424.00	$4,647.68	$9,987.51



The scatter plot of actual vs. predicted prices shows that the neural network performs well for lower priced cars under $100k. For more expensive cars, it tends to predict less accurately. There were many outliers that were predicted to be $50-100k less than their actual prices.

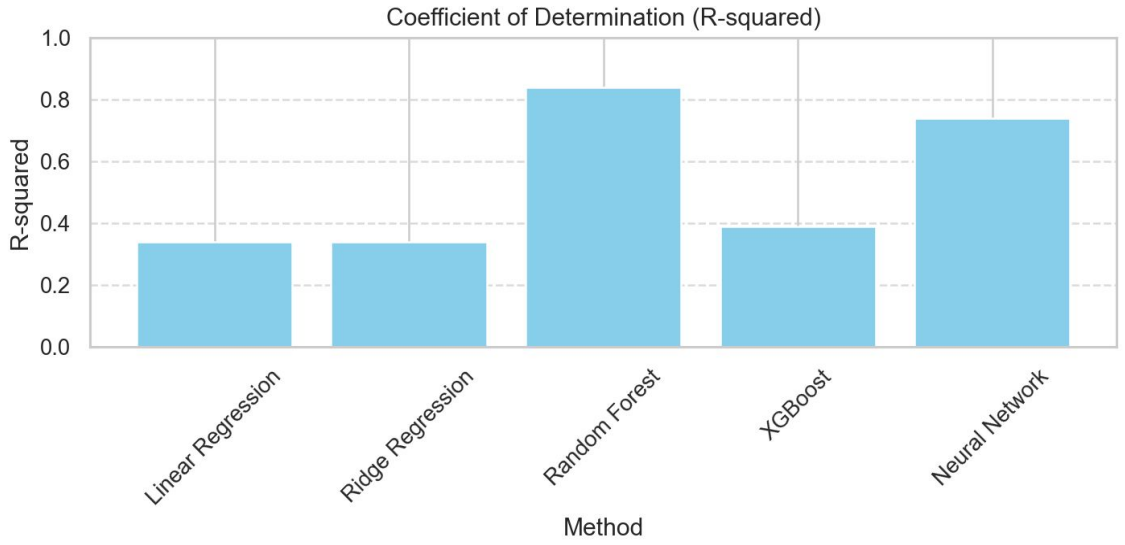
Model Comparison

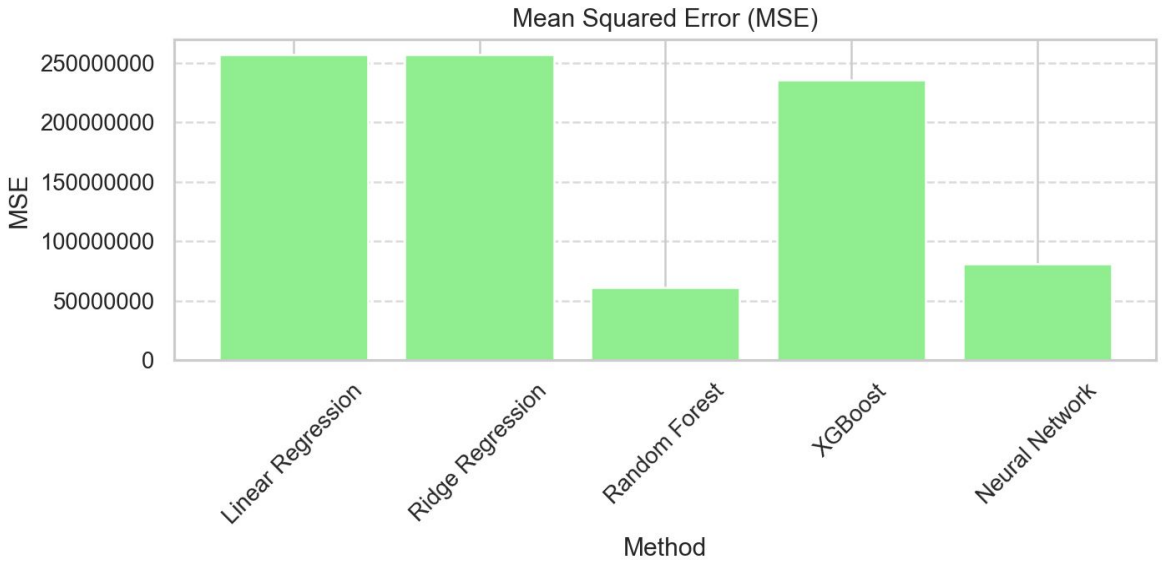
Comparing performance metrics across the 4 models and baseline

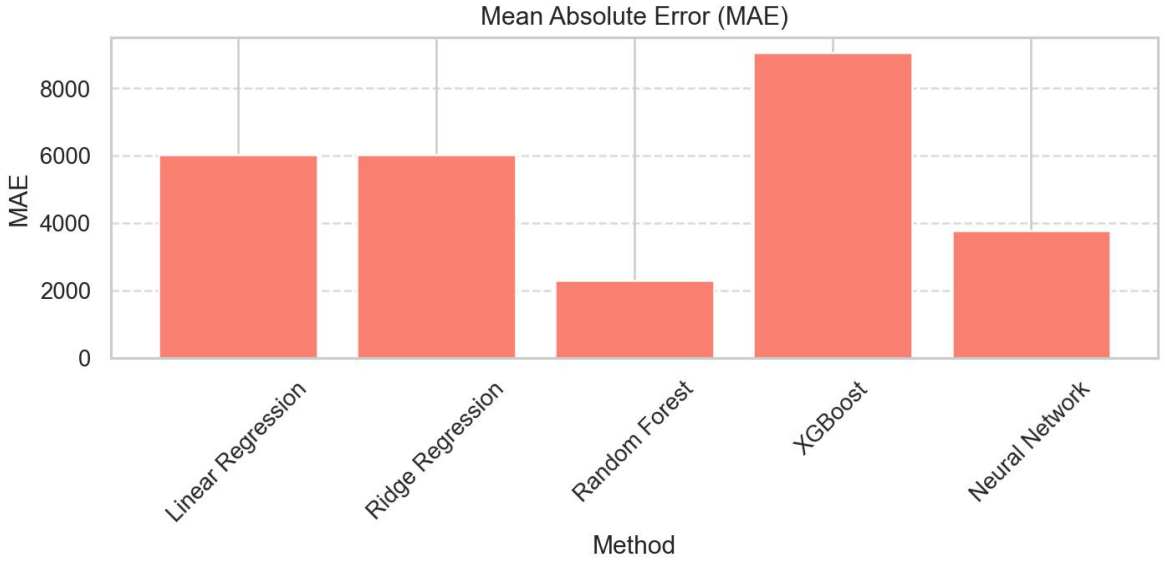
	Linear Regression (baseline)	Ridge Regression	Random Forest	XGBoost	Neural Network
Coefficient of Determination	0.34	0.34	0.84	0.39	0.74
Mean Squared Error (MSE)	$256,894,726.68	$256,892,789.16	$60,702,856.60	$235,901,549.89	$99,750,424.0
Mean Absolute Error (MAE)	$6,029.57	$6029.57	$2,290.23	$9,061.31	$4,647.68
Root Mean Squared Error (RMSE)	$16,027.93	$16027.88	$7,791.20	$15,359.09	$9,987.51

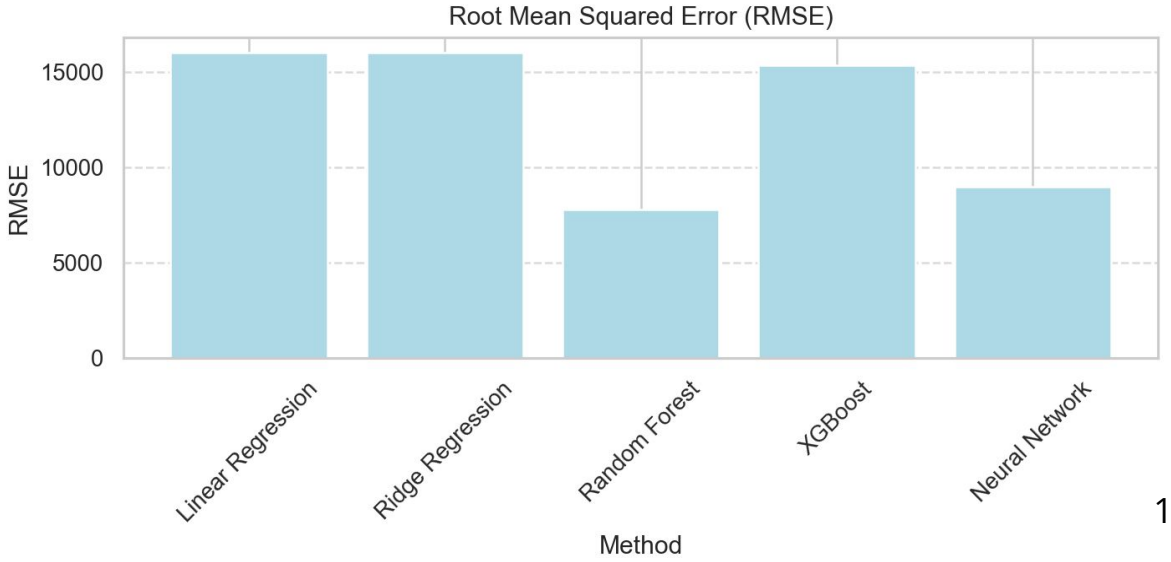
Combined Results Visualization

Bar graphs depicting measures of error and coefficient of determination across all models









Key Takeaways

What We Have Learned

Reducing the Feature Space Decreases Performance

As seen from iterations of each model with different feature sets, reducing the feature set from the base model features had a decrease in performance by approximately 1-2%

Among the Models Mileage Was Most Important

This is in alignment with reality, as many experts recommend checking the mileage as a top factor in determining if the price is a good deal [19]

Random Forest and Neural Network are the best 2 models

The neural network and random forest models performed significantly better than other models (approximately 150% better), with random forest having the highest performance

Normalization Helped Increase Model Accuracy

Normalization techniques (min-max normalization) were tested on the Neural Network models, which resulted in a model accuracy increase of 5%

Challenges & Limitations

What Held Us Back

Outliers in Data Visualization

Outliers made it difficult to visualize and interpret the data as it skewed most plots

Computationally Intensive

Due to the large number of rows and limited device memory, model complexity was limited. As increasing iterations to convergence increased the run time significantly

Hyperparameter Tuning

Grid and random search were attempted, but are computationally intensive, thus a manual tuning approach was taken. In addition, the possible hyperparameters were limited due to some parameters having extremely long training times

Conclusion and Summary

How it Relates to Our Topic

Proposed Model

Random Forest

Random forest had the highest accuracy

Expected Prediction

Accuracy of 84%

Random forest had an accuracy of 84%

Help Buyers Make

Informed Decisions

The model will help buyers get a feel of what the car is worth

References

[1] AnanayMital, “US used cars dataset,” Kaggle, https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset (accessed Mar. 01, 2024).

[2] S. Previl, “Will car prices come down in 2024? Industry experts share their outlook,” Global News, https://globalnews.ca/news/10204235/auto-market-2024-forecast/ (accessed Mar. 14, 2024).

[3] N. Acharya, “Choosing between mean squared error (MSE) and mean absolute error (MAE) in regression: A deep dive,” Medium, https://medium.com/@nirajan.acharya666/choosing-between-mean-squared-error-mse-and-mean-absolute-error-mae-in-regression-a-deep-dive-c16b4eeee603#:~:text=Advantages%20of%20Mean%20Squared%20Error,are%20relatively%20small%20and%20consistent (accessed Mar. 14, 2024).

[4] “What are the advantages and disadvantages of using linear regression for predictive analytics?”, LinkedIn, https://www.linkedin.com/advice/1/what-advantages-disadvantages-using-linear-1e (accessed Mar. 14, 2024).

[5] M. Waseem, “How To Implement Linear Regression for Machine Learning?”, Edureka!, https://www.edureka.co/blog/linear-regression-for-machine-learning/ (accessed Mar. 14, 2024).

[6] “ML – Advantages and Disadvantages of Linear Regression”, GeeksforGeeks, https://www.geeksforgeeks.org/ml-advantages-and-disadvantages-of-linear-regression/ (accessed Mar. 14, 2024).

[7] “What is Linear Regression?”, AWS, https://aws.amazon.com/what-is/linear-regression/ (accessed Mar. 14, 2024).

[8] “What are the advantages and disadvantages of Random Forest?”, AIML.com, https://aiml.com/what-are-the-advantages-and-disadvantages-of-random-forest/ (accessed Mar. 14, 2024).

[9] A. Chakure, “Random Forest Regression in Python Explained”, Builtin, https://builtin.com/data-science/random-forest-python (accessed Mar. 14, 2024).

[10] W. Koehrsen, “Hyperparameter Tuning the Random Forest in Python”, Medium, https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74 (accessed Mar. 14, 2024).

References

[11] K. Marshall, “What is the benefit of ridge regression?”, Deepchecks, https://deepchecks.com/question/what-is-the-benefit-of-ridge-regression/#:~:text=Ridge%20helps%20you%20normalize%20(%E2%80%9Cshrink,sophisticated%20models%20while%20avoiding%20overfitting (accessed Mar. 14, 2024).

[12] “Pros and cons of common Machine Learning algorithms”, Medium, https://medium.com/@eculidean/pros-and-cons-of-common-machine-learning-algorithms-45e05423264f (accessed Mar. 14, 2024).

[13] T. Deori, “Demystifying machine learning challenges: Outliers,” Medium, https://levelup.gitconnected.com/demystifying-machine-learning-challenges-outliers-34aa4f45a1b9 (accessed Mar. 14, 2024).

[14] Rithp, “Evaluating the trade-offs between XGBoost and lightgbm,” Medium, https://medium.com/@rithpansanga/evaluating-the-trade-offs-between-xgboost-and-lightgbm-c1b17fdc4f5e (accessed Mar. 14, 2024).

[15] “What are some of the limitations of XGBoost?: 5 answers from research papers,” SciSpace - Question, https://typeset.io/questions/what-are-some-of-the-limitations-of-xgboost-fniaqa1new (accessed Mar. 14, 2024).

[16] “Introduction to boosted trees,” Introduction to Boosted Trees - xgboost 2.1.0-dev documentation, https://xgboost.readthedocs.io/en/latest/tutorials/model.html (accessed Mar. 14, 2024).

[17] A. Abu-Rmileh, “Be careful when interpreting your features importance in xgboost!,” Medium, https://towardsdatascience.com/be-careful-when-interpreting-your-features-importance-in-xgboost-6e16132588e7 (accessed Mar. 14, 2024).

[18] “Pros and cons of neural networks,” Packt, https://subscription.packtpub.com/book/data/9781788397872/1/ch01lvl1sec27/pros-and-cons-of-neural-networks (accessed Mar. 14, 2024).

[19] T. Ltd. and ThinkInsure, “Best used car buying guide 2023,” How To Buy A Used Car In Ontario | Best Used Car Guide, https://www.thinkinsure.ca/insurance-help-centre/how-to-buy-a-used-car-guide.html (accessed Mar. 14, 2024).

Thank You!