Final Project

LNM Implementation Associated with Used Cars Price in U.S.

Mar 7, 2024

Group 3:

Forough Mofidi
Bonny Mathew
Shaojie Chen
Naoki Tsumoto

Agenda

Context and Objectives

Overall Project Design

Conclusions

The used car market is facing challenges due to its intricate dynamics

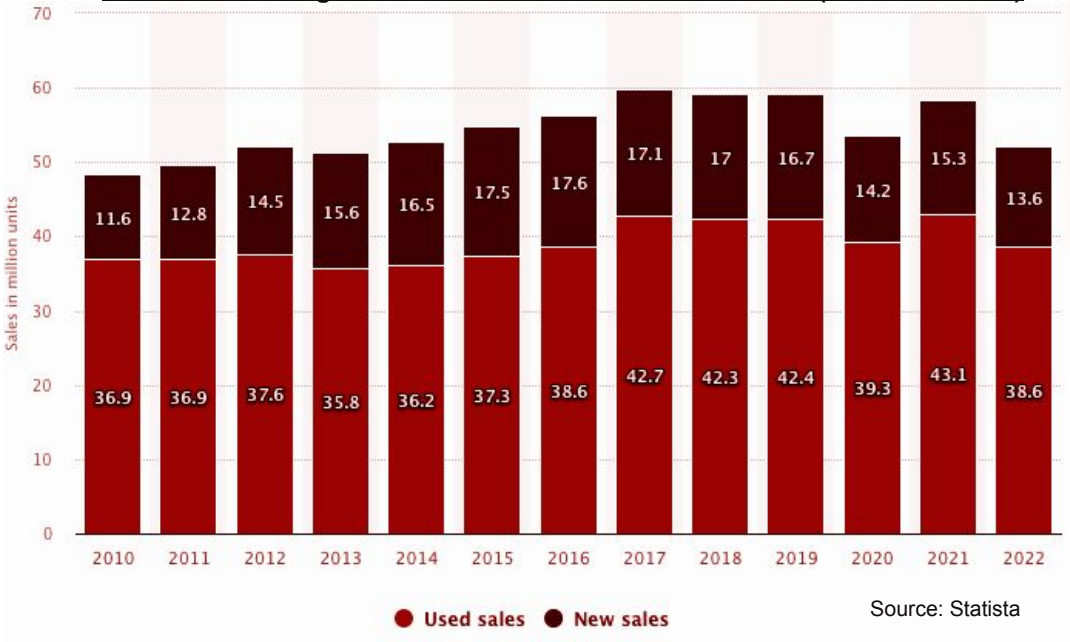
Context:

The used car market exhibits complex dynamics influenced by various factors. Understanding these dynamics is crucial for both consumers and businesses in the automotive industry.

Problem Statement:

Develop a predictive model for used car prices that not only achieves high accuracy but also provides insights into the factors influencing these prices and the causal relationships between them.

New and Used Light Vehicle Sales in US from 2010-2022 (in million units)



Source: Statista

Agenda

Context and Objectives

Overall Project Design

Conclusions

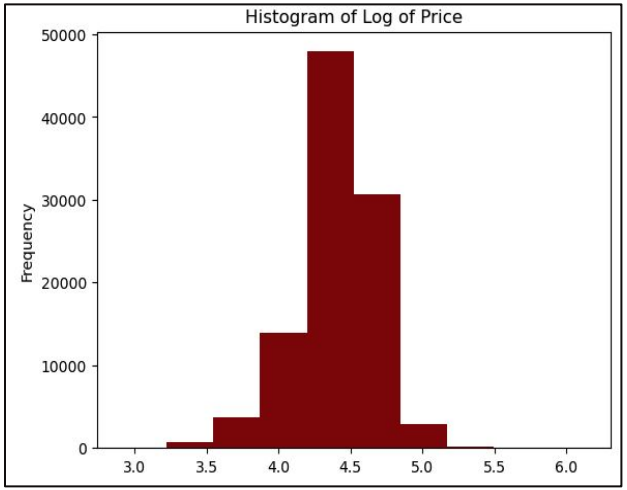
Data Collection & Description

01	Data Source	Used Cars Dataset from Kaggle Website
02	Size of Dataset	3 million observations & 66 variables
03	Data Content	Data obtained by running self made crawler on Cargurus inventory in September 2020
04	Data used for analysis	Random sample of 100k observations
05	Numerical vs Categorical	19 numerical, 47 categorical
06	Sample Variables	Price, daysonmarket, engine_type, fuel_type, body_type, wheel_system etc

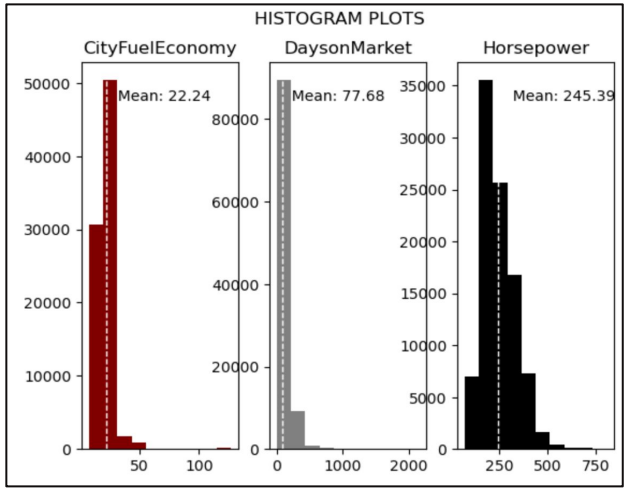
Exploratory Data Analysis (EDA)

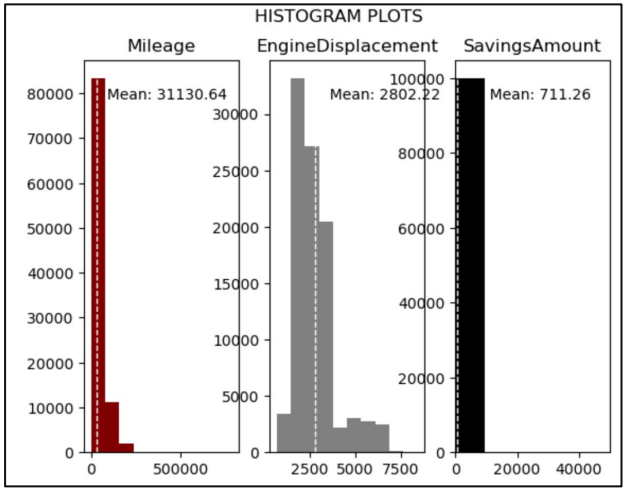
Used a random sample of 100k observations from the 3 millions observations to run EDA and conduct the subsequent modelling.

Histograms of Log of Price (target)



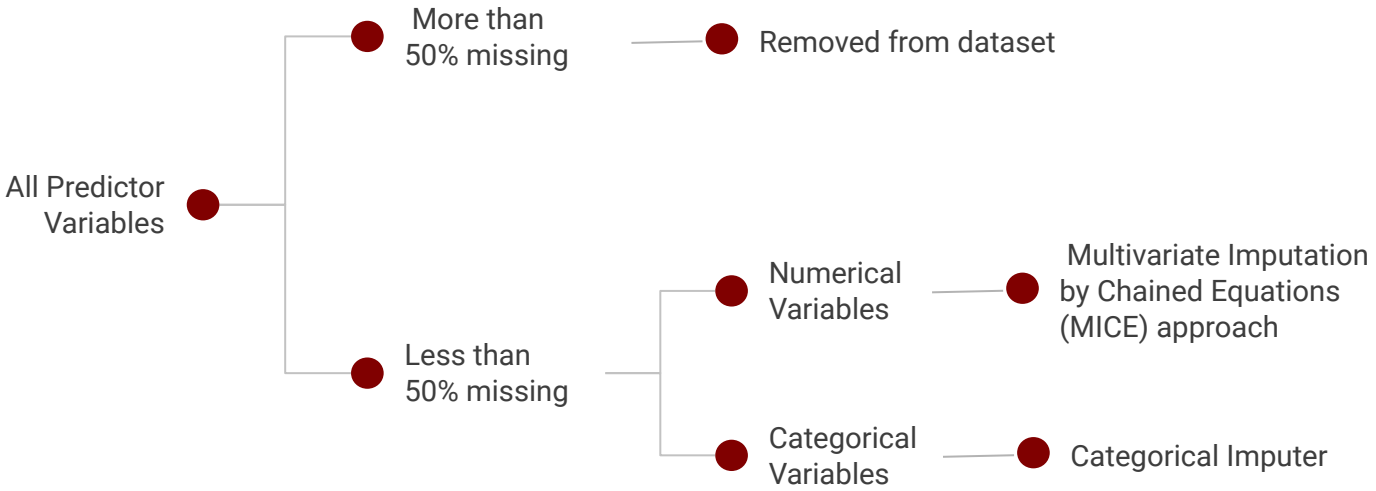
Histograms of some of the main numerical predictors





Data Pre-processing: Missing Value Imputation

After this data pre-processing step 57 out of 65 predictor variables were retained in the dataset



Data Pre-processing: Feature Engineering & Selection

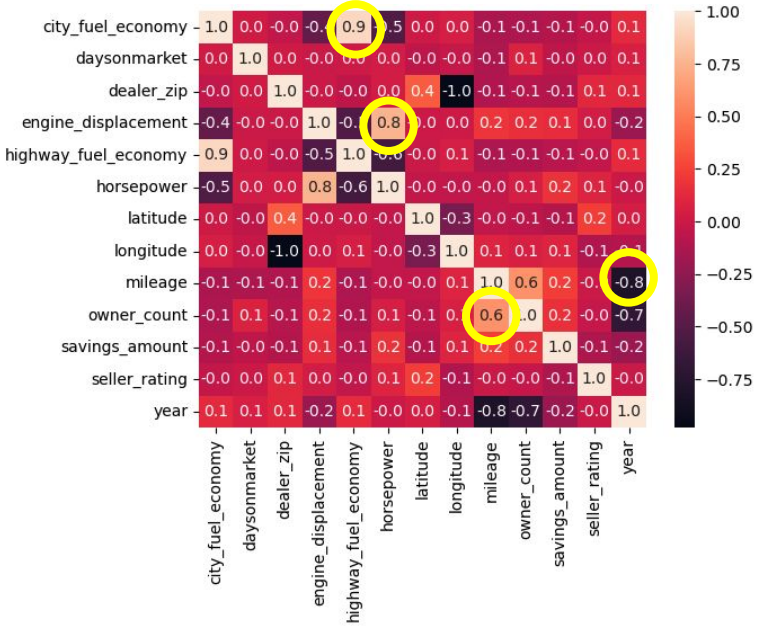
A Pair-wise correlation plot between the numerical variables revealed strong correlations between some of the predictors.

Principal Component Analysis (PCA) was run on some of these highly correlated variables to address redundancy and reduce dimensionality.

Out of the 9 variables which went through PCA we arrived at 4 principal factors

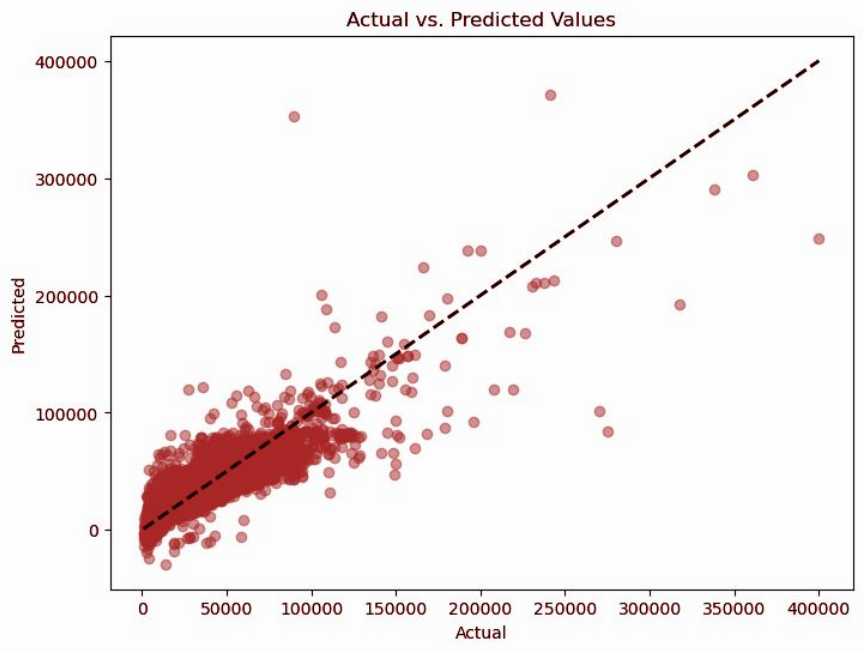
Additionally, some of the variables were removed based on industry knowledge

These principal factors were combined with the remaining variables to arrive at the final dataset for the modelling purpose



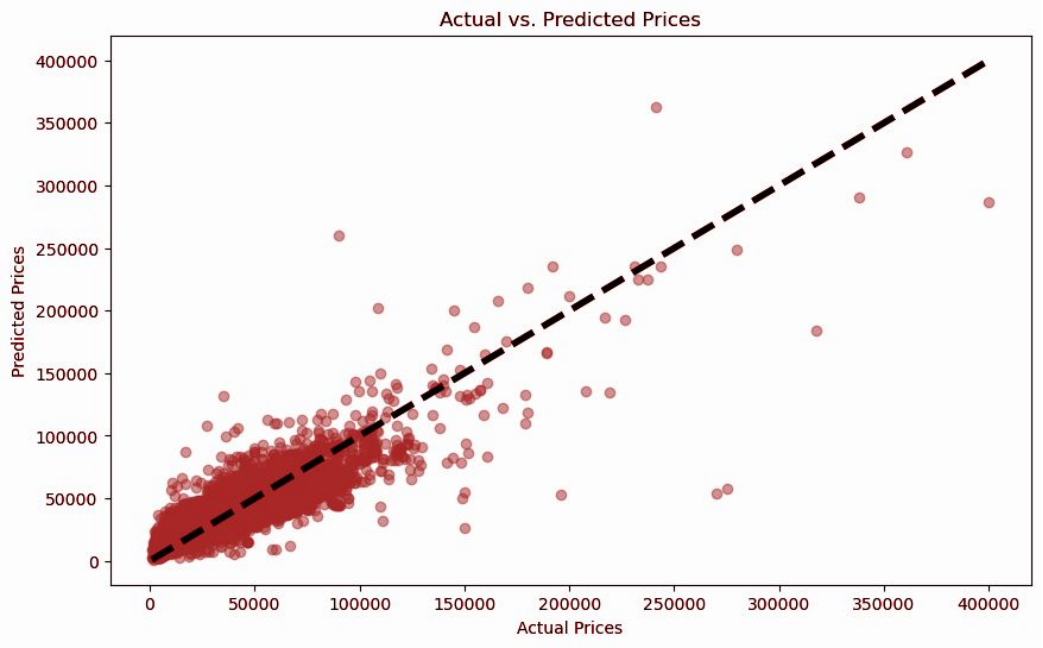
OLS Model

- The car's make name, model name, fuel type, mileage and number of days listed on the platform were used as variables for the linear model.
- Using a train test split of 70:30, the resulting model had an RMSE of 10,001 and R2 score of 0.738.
- Most notably, there is heteroscedasticity observed by the noticeable cone shape and model assumptions are violated here.



GLM Model

- The GLM model with a Gaussian distribution and log link fared better as the log link ensures predictions are positive and to account for the outlier prices that luxury car brands may fetch.
- Using a train test split of 70:30, the resulting model had an RMSE of 9,806 and R2 score of 0.7482.



The Linear Model was unable to adequately explain used car prices, indicating that a non-linear model might be a more suitable choice in this case

Non-Linear Model Approach

- Our literature review revealed that Explainable Boosting Machines and Propensity Score Matching can be a suitable solution for addressing the nuances of the price prediction that was not previously addressed by the regression model.

- Methodology for this project
  - From XAI perspective: **Explainable Boosting Machines**
  - From a Causal Inference perspective: **Propensity Score Matching.**

Explainable Boosting Machines | What is EBM?

EBM

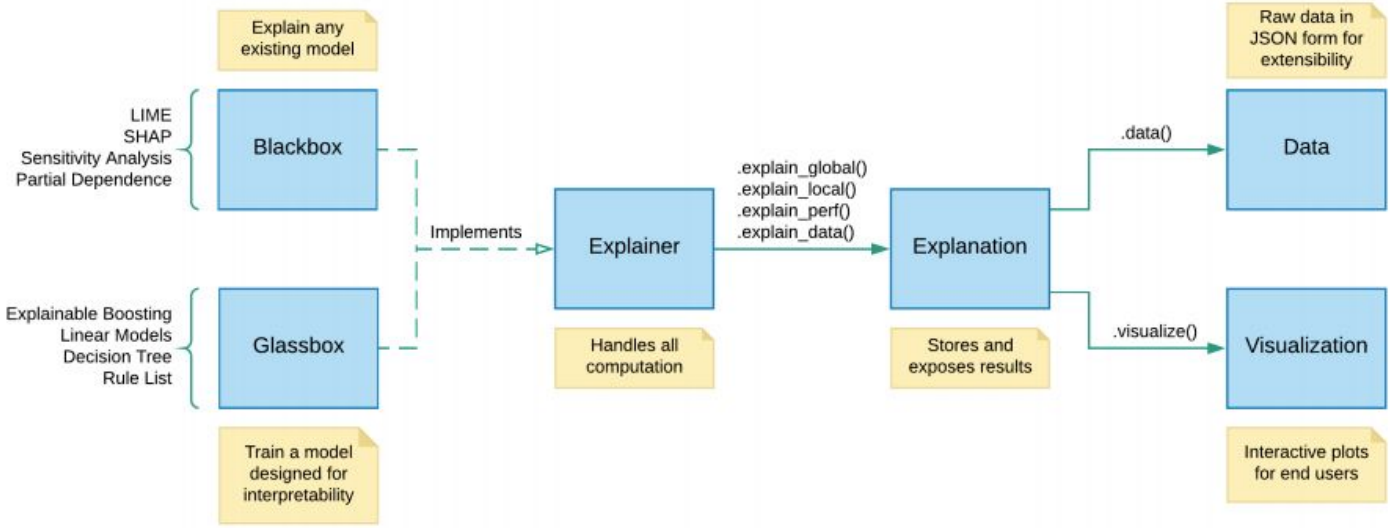
- EBM is a tree-based model
- EBM is a type of generalized additive model (GAM)
- Able to capture non-linearity in the data
- EBM with additive structure has the interpretability of a linear model
- Linear relationship is replaced by several non-linear smooth functions
- EBMs are an improvement on GAM using techniques such as gradient boosting & bagging

Explainable Boosting Machines | Why EBM?

- EBM has light memory usage and has fast predict times
- The boosting procedure is restricted to train on one feature at a time using a very low learning rate
- EBM is a glassbox model, designed to have accuracy comparable to state-of-the-art machine learning methods like Random Forest and Boosted Trees
- EBMs are often as accurate as blackbox models while remaining completely interpretable.

IntepretML: A Unified Framework for Machine Learning Interpretability

(EBMs) are included in a toolkit for Machine Learning Interpretability called InterpretML. It is an open-source package for training interpretable models as well as explaining black-box systems. Within InterpretML, the explainability algorithms are organized into two major sections, i.e., **Glassbox models** and **Blackbox explanations**. This means that this tool can not only explain the decisions of inherently interpretable models but also provide possible reasoning for black-box models.



https://www.kaggle.com/code/parulpandey/explainable-boosting-machines-for-tabular-data

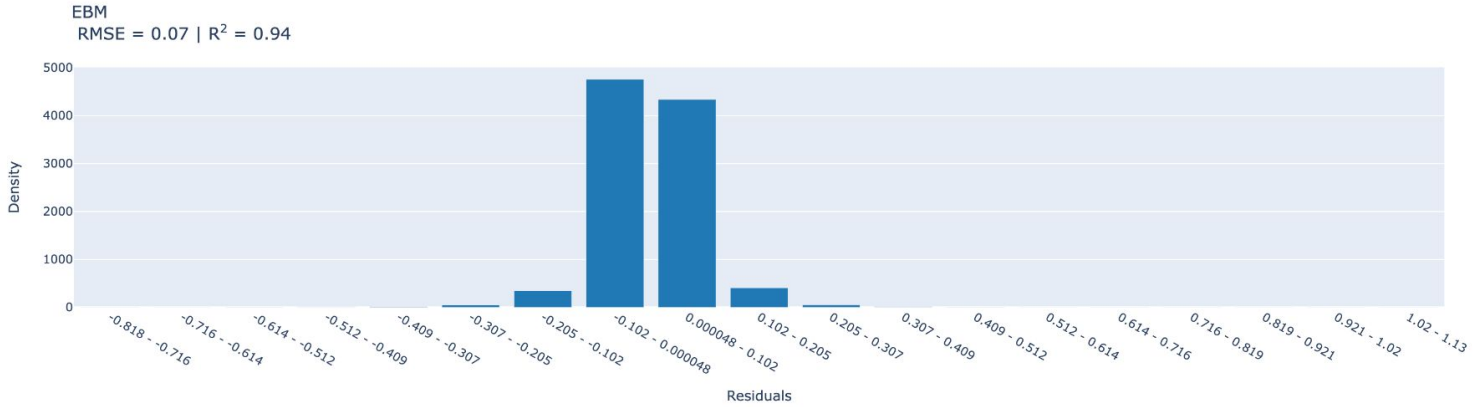
EBM - Model Results

```
[332]: ebm_perf = RegressionPerf(ebm, column_names).explain_perf(X_test, y_test, name='EBM')
       show(ebm_perf)
```

Select Component to Graph

Summary

EBM



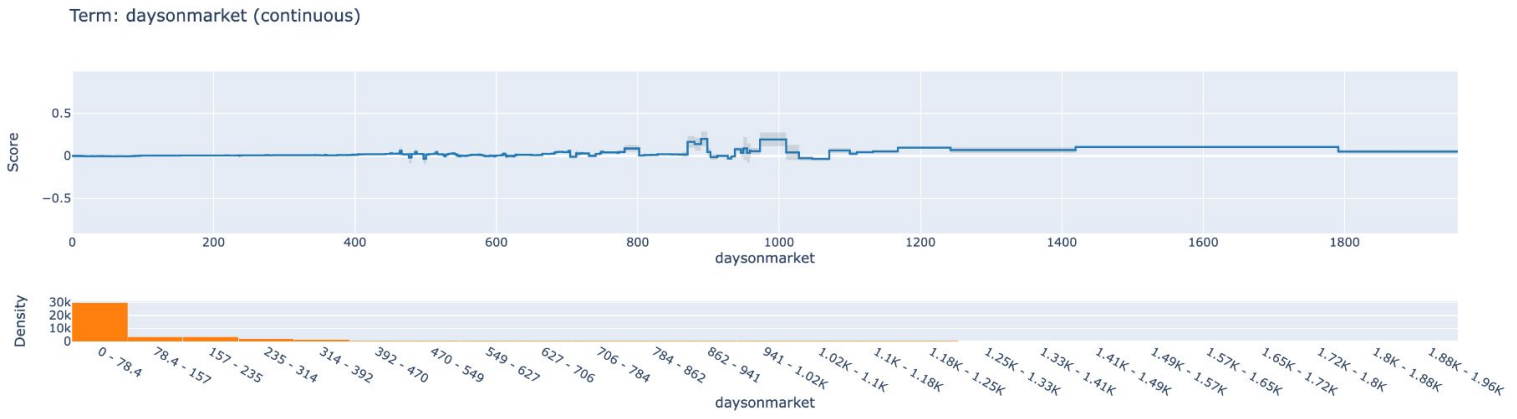
EBM - Model Results

```
[322]: show(ebm.explain_global())
```

Select Component to Graph

0 : Name (daysonmarket) | Type (continuous) | # Unique (688)

ExplainableBoostingRegressor_9



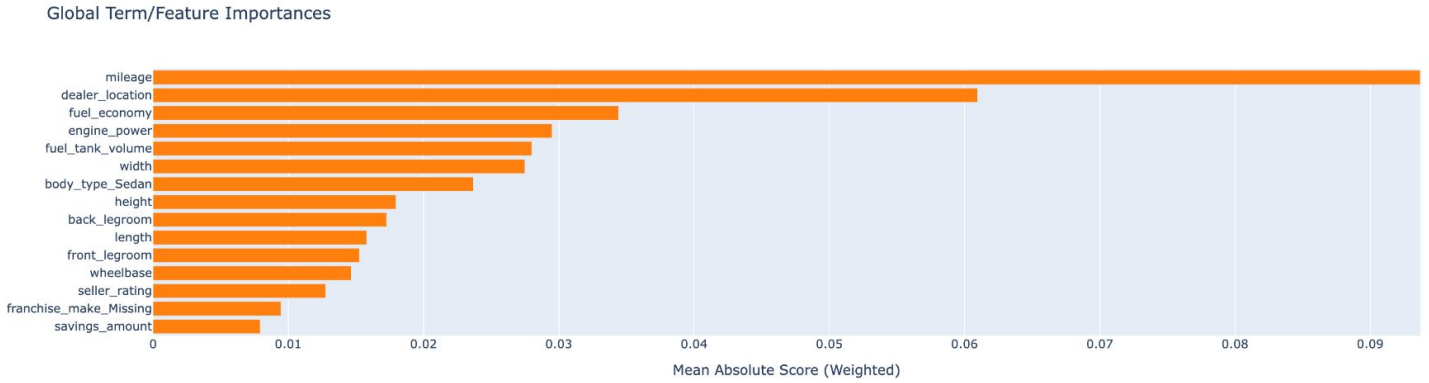
EBM - Model Results

```
show(ebm.explain_global())
```

Select Component to Graph

Summary

ExplainableBoostingRegressor_9



EBM - Model Results

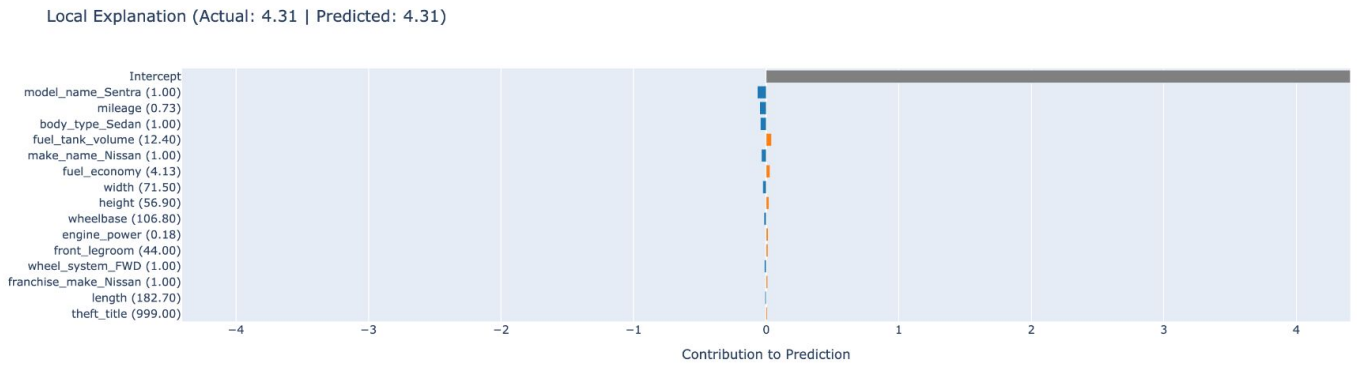
Local Explanations

```
local_explanations = ebm.explain_local(X_test, y_test, name='EBM')
show(local_explanations)
```

Select Component to Graph

0 : Actual (4.306) | Predicted (4.312) | Resid (-0.005)

EBM

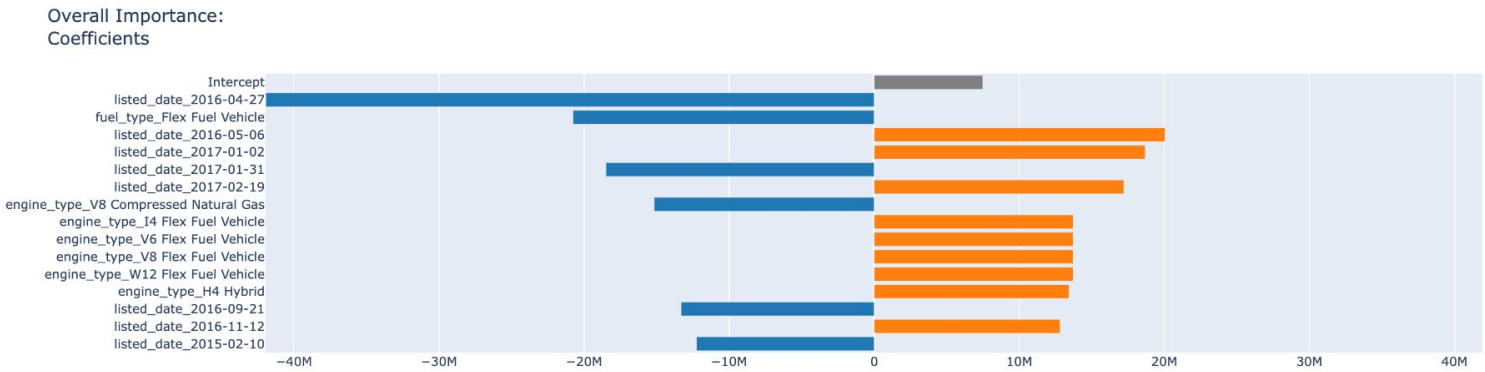


EBM - Model Results

Select Component to Graph

Summary

Linear Regression



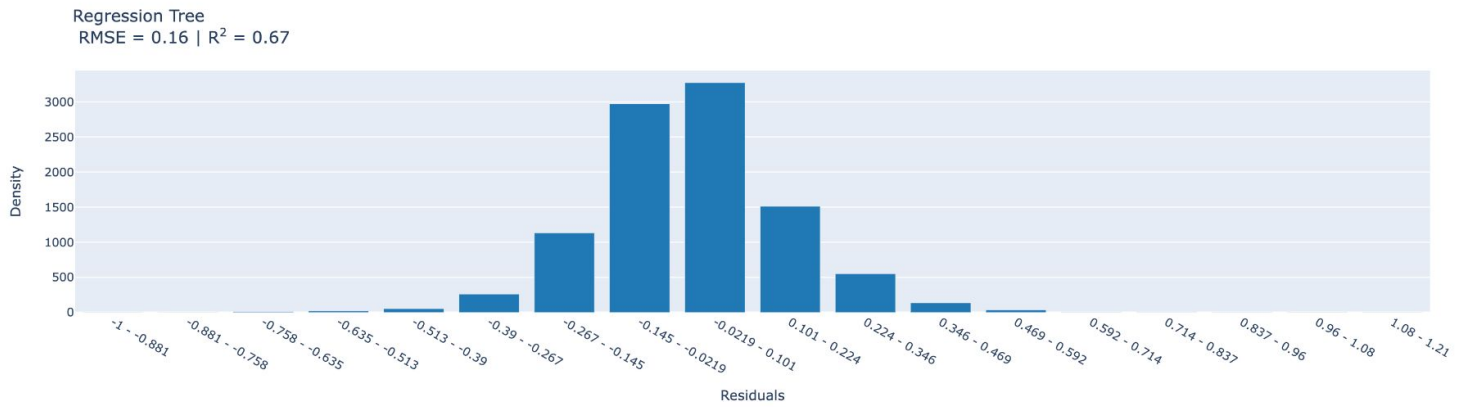
EBM - Model Results

```
rt_perf = RegressionPerf(rt, column_names).explain_perf(X_test, y_test, name='Regression Tree')
show(rt_perf)
```

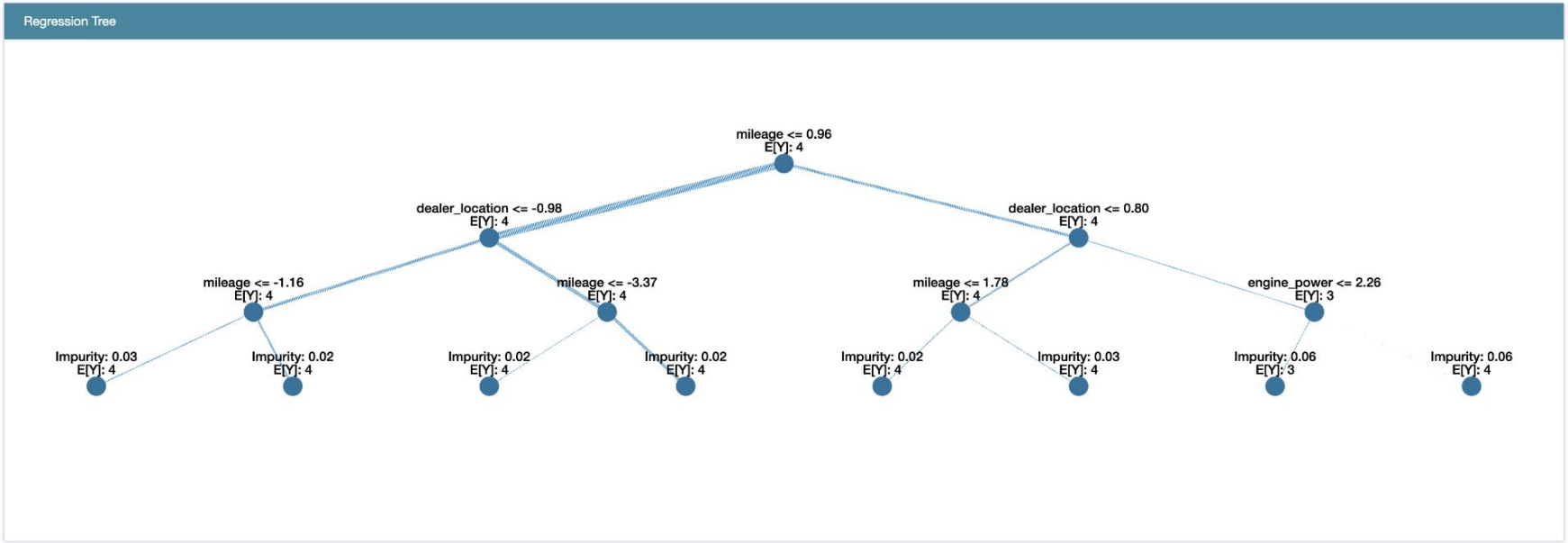
Select Component to Graph

Summary

Regression Tree



EBM - Model Results
Global - Regression Tree



Propensity Score Matching (PSM)

Why PSM?

PSM is a statistical method extensively applied in healthcare and policy analysis to mitigate biases in observational studies. It involves:

- **Estimating Propensity Scores:**
  Calculating the likelihood of receiving the treatment based on observed characteristics, typically using logistic regression.
- **Creating Balanced Datasets:**
  Matching units from treatment and control groups with similar propensity scores to ensure comparable groups, thereby emulating a randomized experiment conditions.
- **Enhancing Accuracy in Treatment Effect Estimation:**
  By adjusting for confounding variables, PSM allows for a more precise estimation of the causal effect of treatments or interventions.

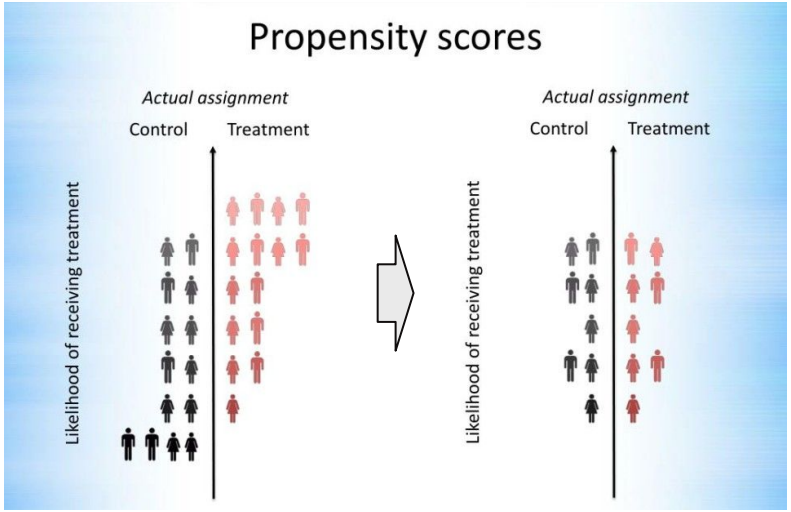


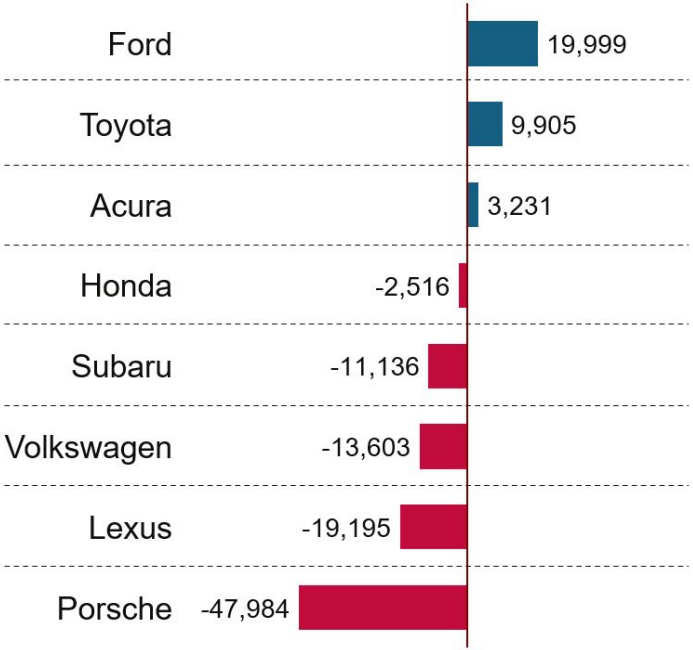
*Image: https://www.linkedin.com/pulse/types-propensity-score-matching-jason-shafrin/

PSM revealed that hybrid vehicles have a negative value in the market but a premium exists for specific brands such as Ford and Toyota

Treatment Effect Estimates

Category	Est.;USD	P > |z|
ATE (Average Treatment Effect)	-24,881	0.039
ATC (Average Treatment effect on the Controls)	-25,049	0.042
ATT (Average Treatment effect on the Treated)	-15,971	0.448

ATE by Brands



Insights

- Overall, it is shown that hybrid vehicles tend to have lower prices in the market.
- However, the premium for hybrid vehicles varies by specific brands, with significant premiums observed for Ford and Toyota.

There are notable challenges and limitations to be mindful of in the datasets and XAI techniques used within the project

Category	Challenges and Limitations
Dataset	• The data is up-to-date as of September 2020, not reflecting the most recent trends • Additionally, due to imperfections in the dataset itself, about 50% of the data was deleted before model construction, meaning the models do not necessarily reflect the entire market data
Explainable Boosting Machines (EBM)	• While EBM is relatively lightweight, training can be time-consuming, especially with large datasets or when the number of features with complex interactions is high • EBM learns individually for each feature, necessitating appropriate regularization and cross-validation to prevent overfitting
Propensity Score Matching (PSM)	• The effectiveness of PSM hinges on the comprehensiveness and accuracy of the observed covariates included in the model. If important factors that influence both the treatment (e.g., being a hybrid vehicle) and the outcome (e.g., market price) are omitted, the resulting estimates may be biased. This is known as "omitted variable bias."

Agenda

Context and Objectives

Overall Project Design

Conclusions

The used car market is difficult to predict due to its complexity, but by utilizing XAI techniques, it is possible to unravel the structure and gain insights

- **Conclusions :**
  - It was found that using non-linear models is necessary to maintain high prediction accuracy in predicting used car prices in the US market due to the complex relationships within the market
  - Particularly, being a hybrid vehicle is an important feature, and it has been confirmed using XAI methods that it brings a premium price of ~20,000 USD for specific brands such as Ford and Toyota
- **Future Study:**
  - The data used in this study is somewhat outdated, being only up to 2020 and thus less influenced by recent environmental shifts towards carbon neutrality
  - Therefore, analyzing the latest data considering current trends, such as electric vehicles, is believed to lead to a more fundamental understanding of the market

Thank You!