DSC 630 Final Project

Dan Wiltse, Dan Zylkowski, Gabby Beinars

Business Objective

Use predictive modeling to determine the value of a used vehicle, based on key features

Data

- 100,000 UK Used Car Data set | Kaggle
- Separate CSV files for each car manufacturer were combined to create a single dataset
  - Audi
  - Mercedes
  - BMW
  - Volkswagen
  - Toyota
  - Hyundai
  - Ford

Data

- 100,000 used cars manufactured from 1997 to 2020
- Ten different variables: make, model, year, price, transmission, mileage, fuel type, tax, mpg, and engine size
- Categorical and numerical data

Exploratory Data Analysis – Data Preparation

- Initial EDA was performed on the datasets individually by vehicle manufacturer prior to combining
- No observations needed to be removed due to multicollinearity
- Preliminary results show moderately strong correlation between price and mileage, indicating that newer and lower mileage vehicles tend to be priced higher
- Various outliers were removed and missing values replaced with the median of that feature, or removed as appropriate

Exploratory Data Analysis – Target Variable

- Price – observed as skewed, requiring transformation

Modeling

- The PyCaret library was used for all steps of the model building process. The data was split into a test set and train set and we initially fit 16 regression models to the data.
- The following models performed best:
  - CatBoost Regressor
  - Random Forest Regressor
  - Extra Trees Regressor
  - Extreme Gradient Boosting
  - Light Gradient Boosting Machine

Modeling – Baseline with no Preprocessing

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	1308.8365	4817397.9441	2189.8866	0.9548	0.0997	0.0733	9.2080
rf	Random Forest Regressor	1252.3543	4858372.3424	2199.2767	0.9545	0.1052	0.0724	13.9640
et	Extra Trees Regressor	1262.4592	4859644.4145	2199.4651	0.9544	0.1061	0.0734	20.9200
xgboost	Extreme Gradient Boosting	1373.2376	5244985.4000	2286.7609	0.9508	0.1059	0.0774	8.7780
lightgbm	Light Gradient Boosting Machine	1512.4595	7014618.8419	2644.8160	0.9342	0.1135	0.0834	0.4040
dt	Decision Tree Regressor	1575.9366	7930510.2871	2809.8298	0.9257	0.1367	0.0925	0.3280
gbr	Gradient Boosting Regressor	2108.2570	12223774.3719	3492.2160	0.8853	0.1512	0.1157	4.9740
br	Bayesian Ridge	1809.5364	13524203.2022	3599.4523	0.8711	0.1349	0.0994	1.3280
ridge	Ridge Regression	1810.6027	13581317.8000	3606.7874	0.8706	0.1349	0.0995	0.1160
lr	Linear Regression	1898.4951	14534325.8000	3737.5970	0.8615	0.1423	0.1054	1.1700
omp	Orthogonal Matching Pursuit	2457.4662	24271435.8097	4805.7233	0.7682	0.1763	0.1321	0.1260
huber	Huber Regressor	3482.4760	28711508.6242	5356.7605	0.7303	0.2917	0.2150	6.6520
ada	AdaBoost Regressor	3718.6857	30855291.1452	5549.2842	0.7105	0.2684	0.2303	9.9460
knn	K Neighbors Regressor	5209.7134	56775023.3181	7534.4157	0.4663	0.3949	0.3301	1.2640
en	Elastic Net	5849.0501	77059011.2000	8776.6531	0.2762	0.4275	0.3696	0.1520
lasso	Lasso Regression	5892.8269	78372803.2000	8851.2072	0.2638	0.4293	0.3719	0.5860

Modeling – Baseline after Preprocessing

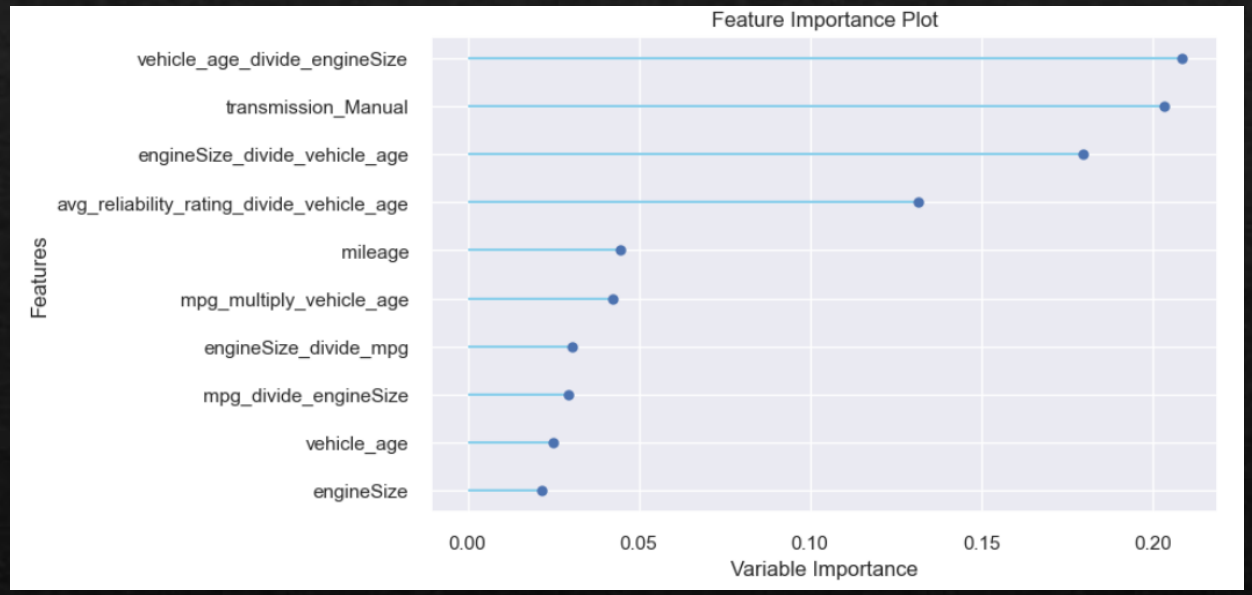
	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	1300.7116	4461113.2859	2109.3782	0.9584	0.0989	0.0728	9.3100
et	Extra Trees Regressor	1266.1621	4500797.6848	2118.4954	0.9580	0.1050	0.0737	19.7520
rf	Random Forest Regressor	1252.9935	4635934.6018	2150.1402	0.9568	0.1044	0.0725	15.4960
xgboost	Extreme Gradient Boosting	1371.1532	4911842.9000	2214.9315	0.9543	0.1051	0.0771	8.8160
lightgbm	Light Gradient Boosting Machine	1502.5054	6661726.4450	2578.2893	0.9379	0.1126	0.0829	0.3620
lr	Linear Regression	1853.6501	9752153.8000	3121.7710	0.9094	0.1331	0.1015	0.4520

Feature Engineering

- PyCaret was used to engineer features.
  - The model variable was excluded from feature engineering.
  - Two features were added prior to feature engineering: avg_reliability_rating and vehicle_age.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
rf	Random Forest Regressor	1428.2621	6267605.4112	2501.4186	0.9416	0.1146	0.0805	8.0940
catboost	CatBoost Regressor	1486.3524	6364225.9118	2521.7086	0.9407	0.1111	0.0814	9.8220
et	Extra Trees Regressor	1484.9404	6764310.7967	2598.2756	0.9370	0.1200	0.0841	6.3540

Feature Importance of Engineered Features



- The following features were added after feature engineering:
  - engineSize_divide_vehicle_age, mpg_multiply_vehicle_age, avg_reliability_rating_divide_vehicle_age, mpg_divide_engineSize

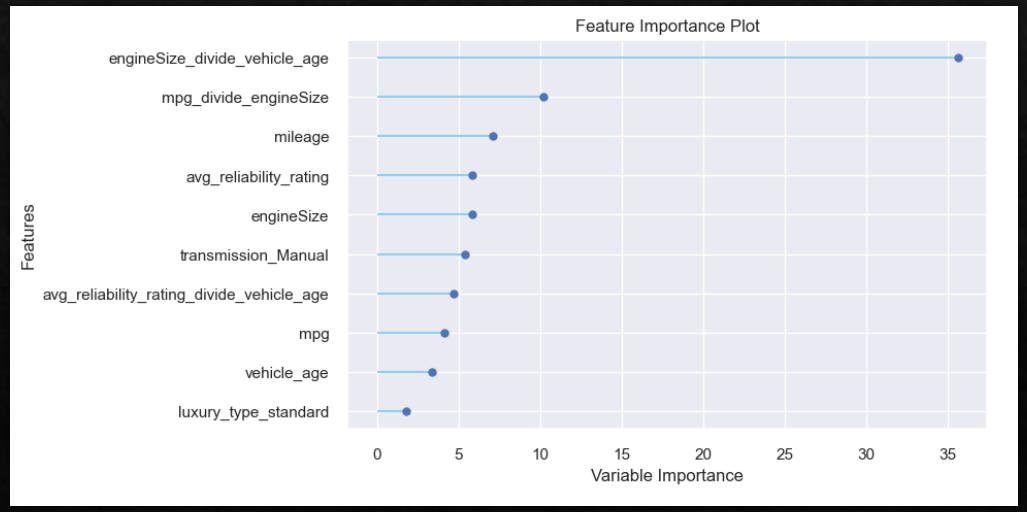
Enhanced Results

- Incremental improvement across almost all model performance metrics
- Catboost still top performing model after feature enhancement

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	1286.7733	4430086.6371	2101.9463	0.9587	0.0979	0.0719	7.8580
et	Extra Trees Regressor	1259.6119	4501962.9718	2118.5974	0.9580	0.1037	0.0729	16.2540
rf	Random Forest Regressor	1239.7949	4550372.6438	2129.7354	0.9576	0.1029	0.0715	13.0280
xgboost	Extreme Gradient Boosting	1350.4506	4888782.7000	2208.7796	0.9544	0.1034	0.0756	6.8100
knn	K Neighbors Regressor	1361.9243	6041224.7782	2452.4711	0.9436	0.1102	0.0772	3.5420
lightgbm	Light Gradient Boosting Machine	1479.3297	6500076.1964	2545.4297	0.9394	0.1108	0.0816	0.7500

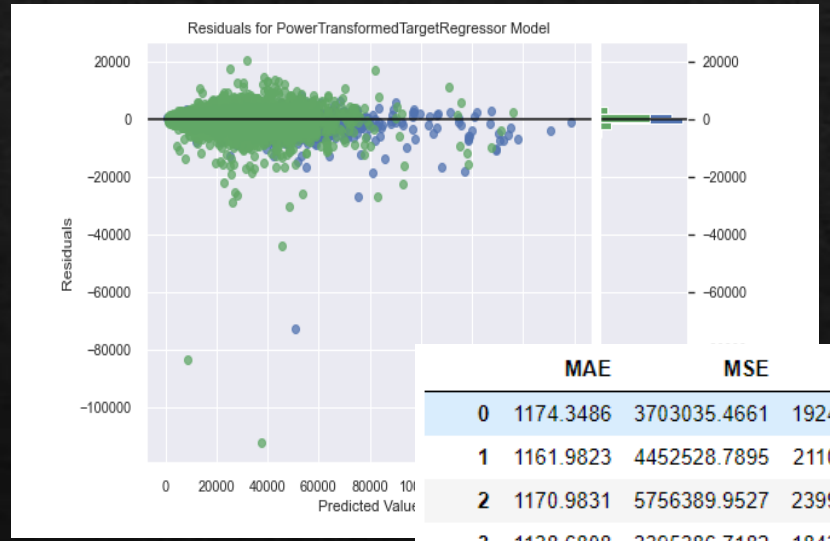
Feature Importance

- Plot of feature importance for top performing model (Catboost)
- Three of top four variables were derived from the data or added to dataset after baseline data was loaded, showing importance of feature selection in model performance



Ensemble Modeling

- Blending ensemble technique used to create multiple models and average the individual predictions to form a final prediction
- Blended data from top three performing models (Catboost, ExtraTrees and Random Forest
- Had highest R squared value of all modeling techniques (.9643)



	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	1174.3486	3703035.4661	1924.3273	0.9663	0.0975	0.0675
1	1161.9823	4452528.7895	2110.1016	0.9575	0.0938	0.0662
2	1170.9831	5756389.9527	2399.2478	0.9410	0.0996	0.0678
3	1138.6808	3395386.7182	1842.6575	0.9681	0.0897	0.0656
4	1220.2537	3868530.6795	1966.8581	0.9673	0.0939	0.0684
5	1149.1336	3364178.4687	1834.1697	0.9696	0.0938	0.0671
6	1150.9393	3383685.8459	1839.4798	0.9686	0.0954	0.0666
7	1116.3978	2980912.0555	1726.5318	0.9717	0.0904	0.0651
8	1173.3733	3672058.6058	1916.2616	0.9657	0.0919	0.0664
9	1176.4441	3507698.1653	1872.8850	0.9676	0.0937	0.0663
Mean	1163.2537	3808440.4747	1943.2520	0.9643	0.0940	0.0667
SD	26.1919	746078.1354	179.4773	0.0086	0.0029	0.0010

Final Results

Model Type	Model Name	R^2	RMSLE
Baseline	CatBoost Regressor	0.9548	0.0997
Baseline	Linear Regression	0.8629	0.1036
Enhanced	CatBoost Regressor	0.9587	0.0979
Enhanced	Linear Regression	0.9129	0.1292
Ensemble	Boosting	0.9592	0.1022
Ensemble	Bagging	0.9603	0.099
Ensemble	Blending	0.9643	0.094

- Blended ensemble model had best performance, followed by Catboost Regressor after feature engineering
- Linear regression showed most improvement from baseline to enhanced modeling

Conclusion

- Exploratory Data Analysis along with data blending and feature engineering improved results from baseline dataset
- Ensemble modeling, especially blending, produced highest accuracy
- Deviation from expected mpg and price for newer vehicles may have impacted model performance
- Additional supplemental information around specific models would help increase predictive ability of future model updates.