NEW YORK UNIVERSITY

Car Sales Prediction

- Lyuang Fu
- Junge Zhang
- Yuanxi Sun
- Yichao Shen

NEW YORK UNIVERSITY

Business Background

- With increasing competition in the auto market, many elements will decide the sales of a car in the market
- Predicted car sales is an important reference for company before a new car make put in the market
- Our team aims to design an algorithm to predict the sales of car models based on different cars' configurations(different features)which help the company determine whether the upcoming car will have a good marketing performance

NEW YORK UNIVERSITY

Data Processing

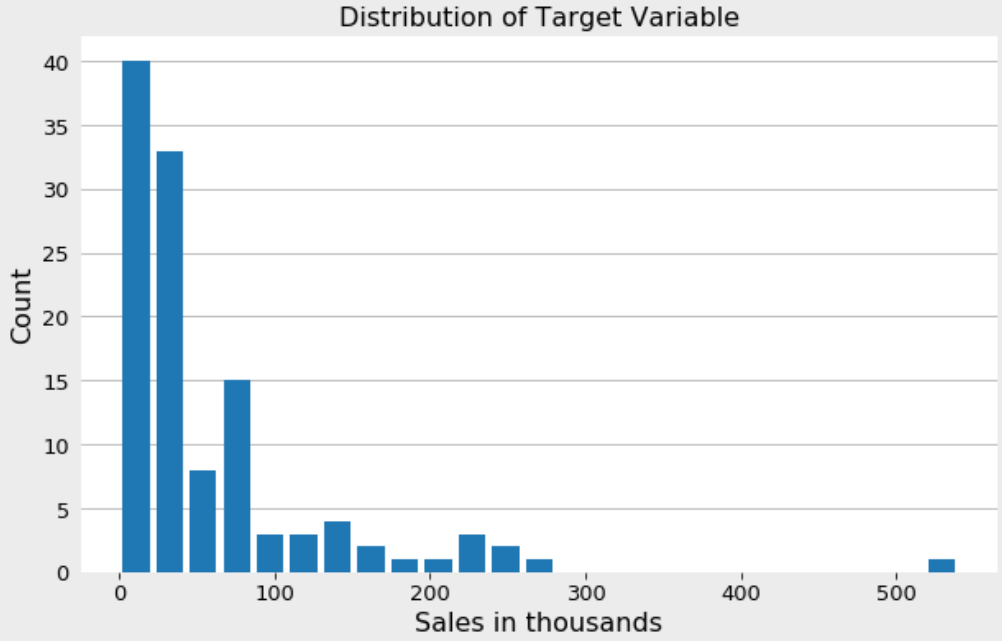
- Data features introduction

  Originally with 15 columns and sales in price as the target variable

- Drop missing data and outliners

  Drop data which are “.”

  Drop data whose values are too much out of range



NEW YORK UNIVERSITY

Data Processing

- Convert text data into numbers

  e.g. convert “Manufacture” to binary values;

  New columns of Acura, Audi and etc.

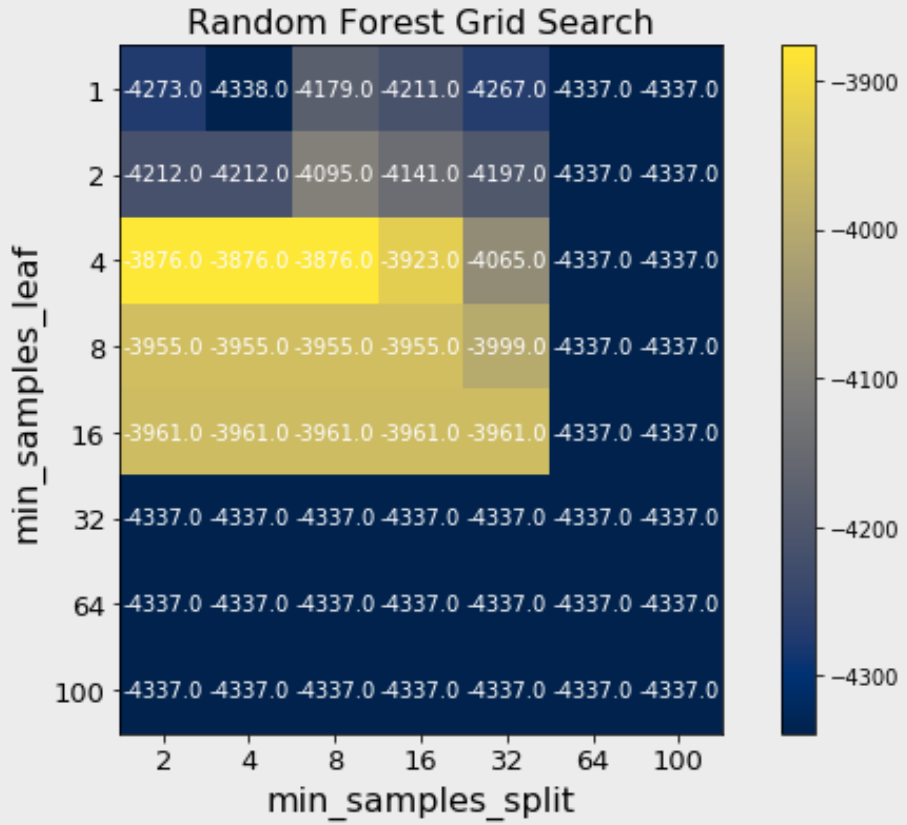
  {Drop business unrelated columns

  e.g. drop the “model” column, “4-year resale value” column and “latest launch” column

NEW YORK UNIVERSITY

Methodology

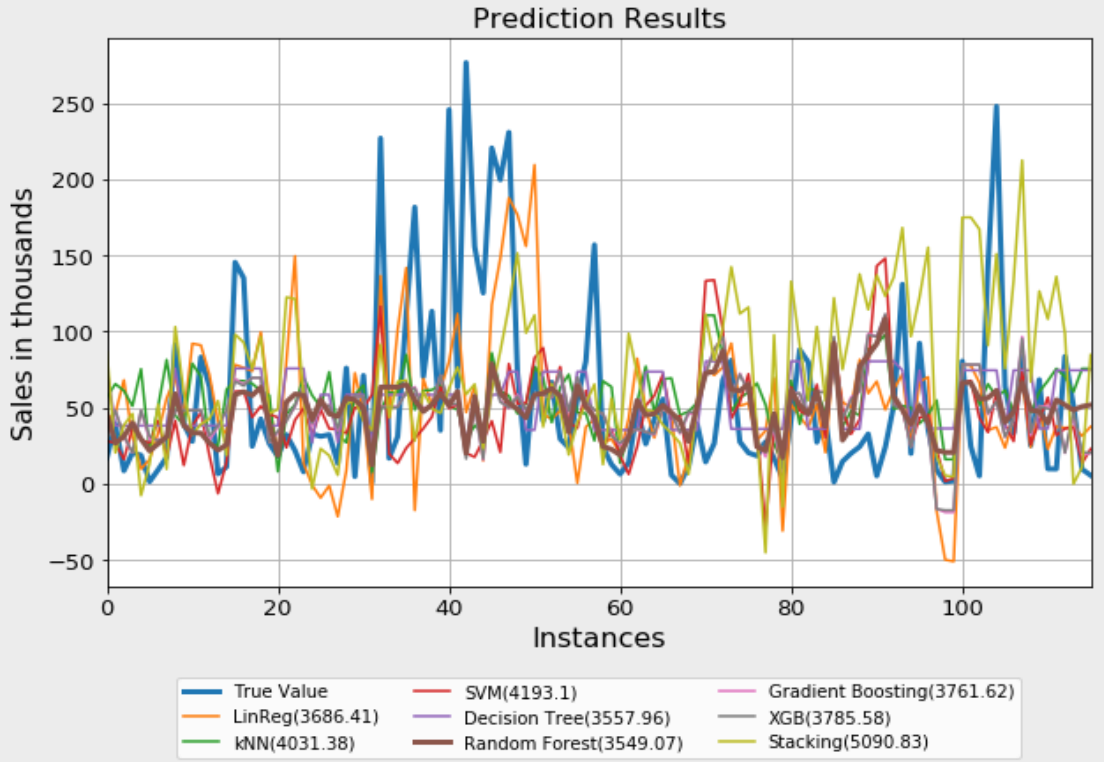
- Multiple regression algorithms are applied under the cross validation grid search to find the best model



NEW YORK UNIVERSITY

Result

- The final determined model is Random Forest with hyperparameter of “max_depth” of 3, “min_samples_leaf” of 4 and “min_samples split” of 2
- The final MSE is 3549.07

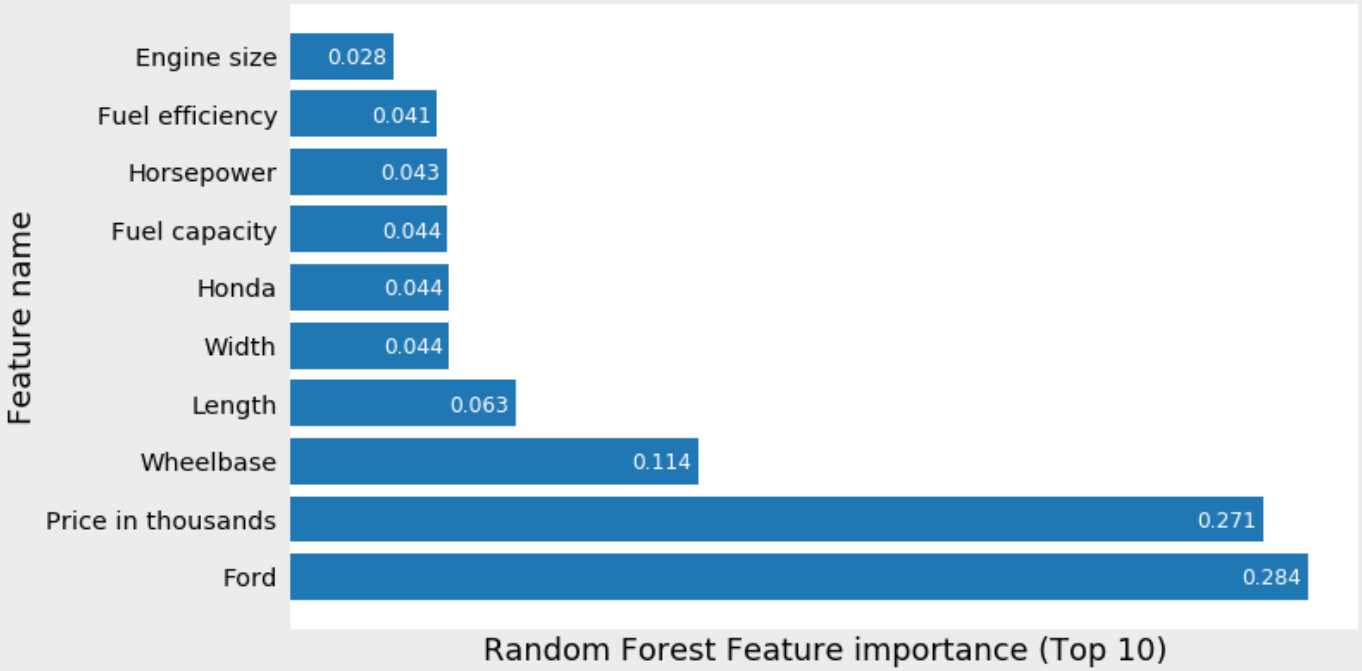


NEW YORK UNIVERSITY

Analysis

- Pros of the model

Feature importance can be used for feature understanding



NEW YORK UNIVERSITY

Analysis

- Cons of the model

The model does not perform well for sales over 200 thousand. However, since there are few such instances, we believe it is because the sample size is too small

	Manufacturer	Model	Sales in thousands	4-year resale value	Price in thousands	Engine size	Horsepower	Wheelbase	Width	Length	Curb weight	Fuel capacity	Fuel efficiency	Latest Launch	Passenger
40	Dodge	Ram Pickup	227.061	15.060	19.460	5.2	230	138.7	79.3	224.2	4.470	26.0	17	16224	False
49	Ford	Taurus	245.815	10.055	17.885	3.0	155	108.5	73.0	197.6	3.368	16.0	24	16789	True
52	Ford	Explorer	276.747	16.640	31.930	4.0	210	111.6	70.2	190.7	3.876	21.0	19	16185	False
55	Ford	Ranger	220.650	7.850	12.050	2.5	119	117.5	69.4	200.7	3.086	20.0	23	16084	False
56	Ford	F-Series	540.561	15.075	26.935	4.6	220	138.5	79.1	224.5	4.241	25.1	18	16298	False
58	Honda	Accord	230.902	13.210	15.350	2.3	135	106.9	70.3	188.8	2.932	17.1	27	16210	True
137	Toyota	Camry	247.994	13.245	17.518	2.2	133	105.2	70.1	188.5	2.998	18.5	27	16710	True

Also the model is not easy to understand and edit without specific knowledge learned

NEW YORK UNIVERSITY

Conclusion

- The chosen model is a better fit to predict the sales of the car and it can be the key reference for auto companies
- Still need new and fresh data to update the model and improve the performance

NEW YORK UNIVERSITY

Reference

- Kaggle. *Car_sales.* [online] Available at: https://www.kaggle.com/sunyuanxi/car-sales-python/data [Accessed 16 Nov. 2018].

NEW YORK UNIVERSITY

Appendix

- Code reference: Kaggle collaborative kernel. https://www.kaggle.com/sunyuanxi/car-sales-python.