CAPSTONE PROJECT
USED CAR PRICE PREDICTION

MACHINE LEARNING/REGRESSION

Madoka Fujii

KEY TAKEAWAYS

- Uncertainty accuracy of price of used car in India while huge demand
- Objective is making a pricing model
- Gradient Boost is the best model in the models explored
- Next step, possible to try analyzing picture and video for prediction.

PROBLEM DEFINITION

Used car market is rapidly growing in India

The price of used car keeps rising dramatically

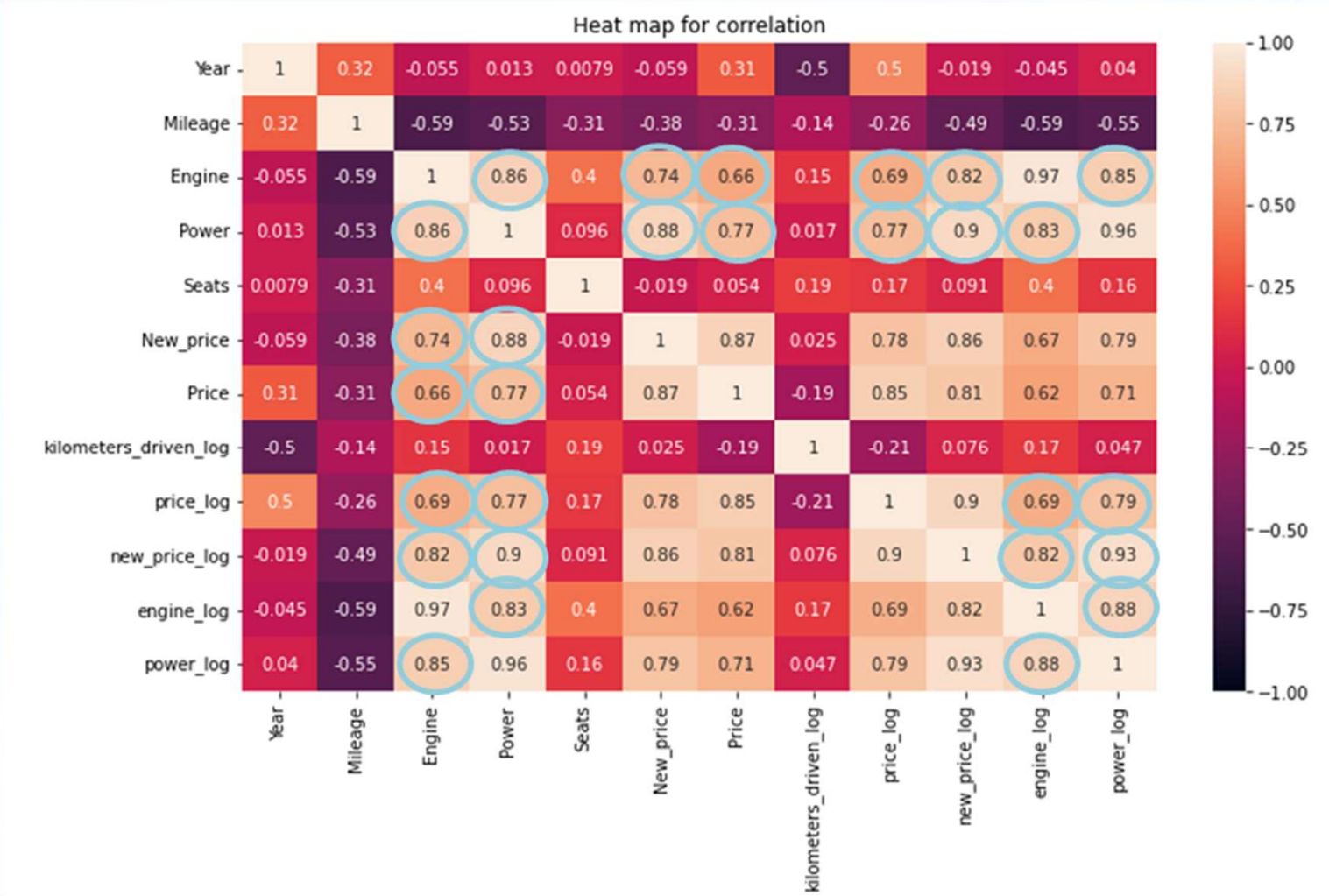
Pricing is hard in terms of accuracy and time

PROBLEM TO SOLVE

- Need to figure out what is the impact to price of used car.
- Can we use machine learning to predict fair price?
- Can the pricing model predict price precisely and appropriately?

EXPLORATORY DATA ANALYSIS

- Engine and Power are highly correlated
- The two correlated with Price and price_log as well

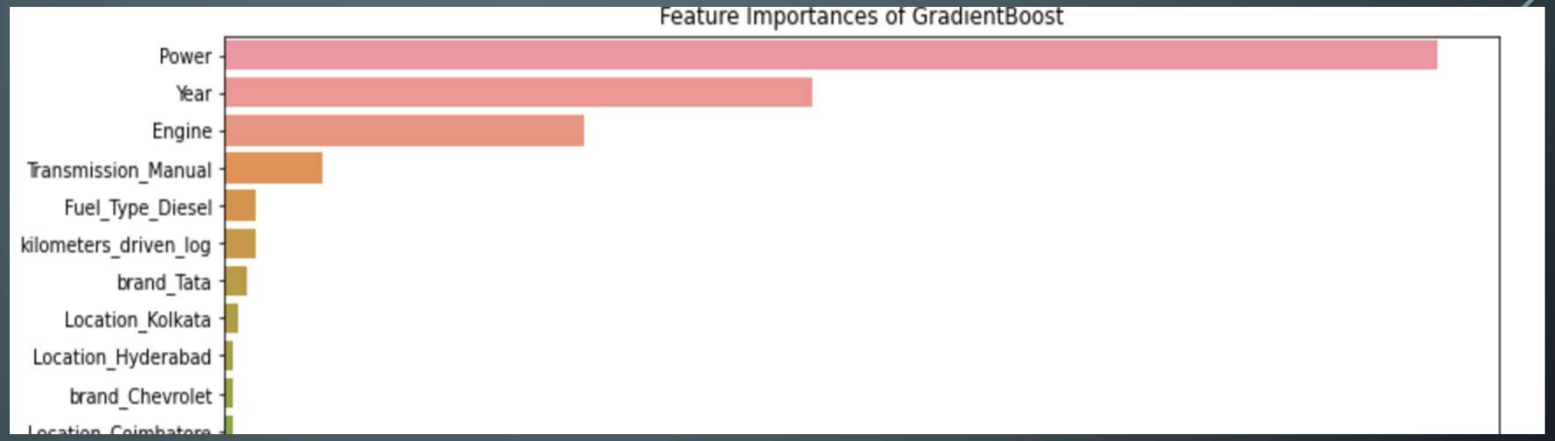


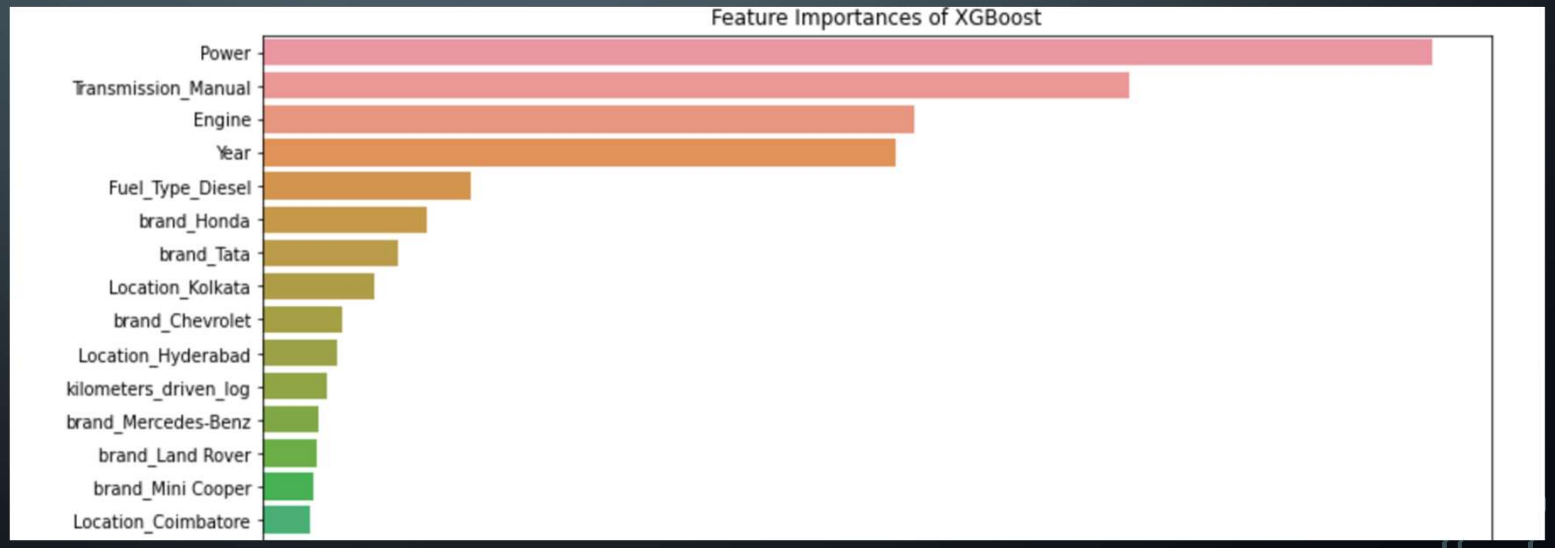
COMPARISON OF MODELS / APPROACH FOR THE SOLUTION

	Model	Train_r2	Test_r2	Train_RMSE	Test_RMSE
0	Decision Tree	0.999997	0.789326	0.020693	5.115459
1	Ridge	0.852567	0.860248	4.289918	4.166371
2	Lasso	0.854212	0.862063	4.265922	4.139239
3	Random Forest	0.972382	0.873334	1.856723	3.966523
4	Turned Decision Tree	0.818296	0.781772	4.762479	5.206360
5	Turned Random Forest	0.926568	0.849099	3.027564	4.329382
6	Tuned_KNN	0.934302	0.827185	2.863694	4.633081
7	XGBoost	0.912270	0.892177	3.309223	3.659606
8	AdaBoost	0.745015	0.690030	5.641680	6.204959
9	GradientBoost	0.922227	0.895283	3.115763	3.606513
10	CatBoost	0.872354	0.865623	3.991676	4.085473

FEATURE IMPORTANCE / KEY FINDINGS & INSIGHTS

- XGBoost is highly sensitive to outliers which makes the order of feature importance becomes different from other models





MODEL SOLUTION

- Gradient Boost can predict price the most accurately and have superior explanatory ability data in the models
- It is robust to outliers. It does not need to do any extra process to evaluate and drop outliers

Model	Train_R2	Test_R2	Train_RMSE	Test_RMSE
GradientBoost	0.92	**0.89**	3.11	**3.60**

RECOMMENDATIONS FOR IMPLEMENTATION

- Need to improve the accuracy of the model that still have remaining error
- Increase the data amount so that it can do more accurate prediction and more profitable

WHAT ARE THE KEY RISKS AND CHALLENGES?

- Accuracy of data: During many detailed processes to create model, it may cause misinterpret data
- The pricing model may not right

NEXT STEP

- Further precise valuation of used car, taking picture or video of the condition (scratches, accidents, or repairs) will be effective. Deep Learning can analyze and predict it. The conditions significantly impact to the car's price.

THANK YOU!
QUESTION?