Final Presentation
Of
Used Car Trading Price Prediction

Tongtong Zhao

Brown University DSI

Dec 10^{th} 2021

https://github.com/Christy9615/DATA1030FinalProj

BROWN

I. Project Recap

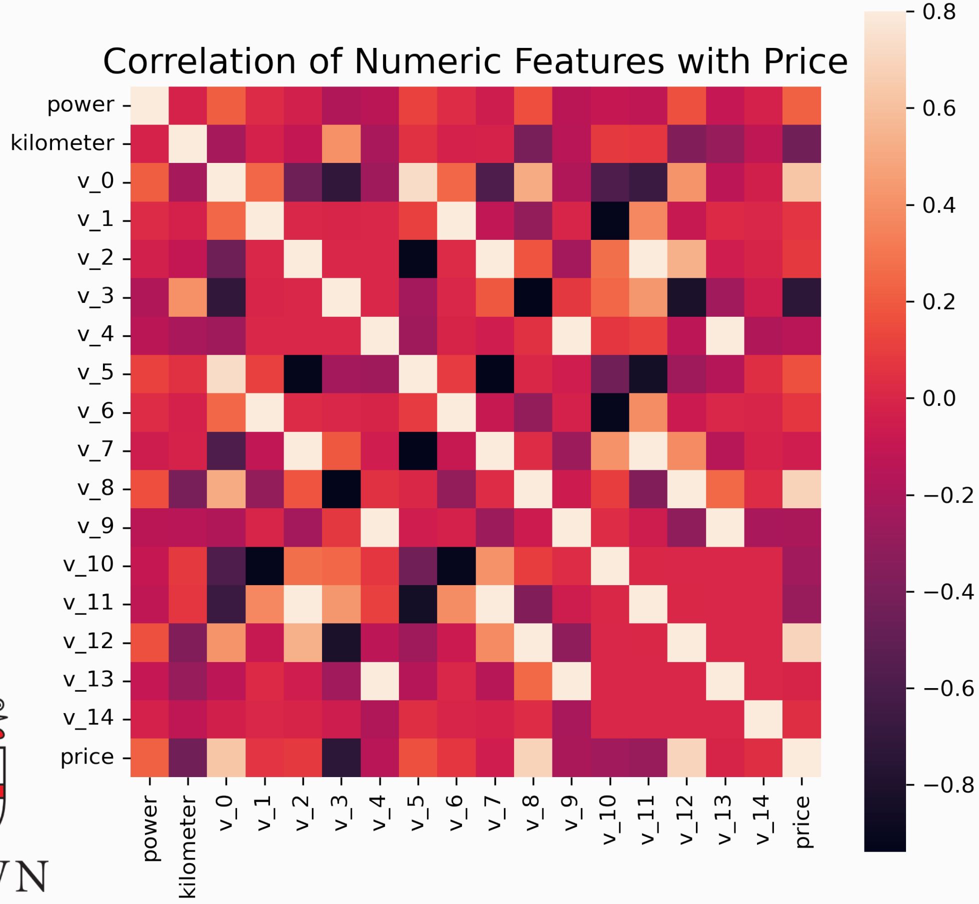
- Dataset Overview --- Used Car Trading Price Prediction
- Project Goal: Predict the trading price of used cars in the test dataset (Regression)

Categorical	seller/ OfferType/ bodyType/ fuelType/ gearbox/ notRepairedDamage regDate/ creatDate/ regionCode/ model/ brand
Continuous	'power', 'kilometer', 'v_0', 'v_1', 'v_2', 'v_3', 'v_4', 'v_5', 'v_6', 'v_7', 'v_8', 'v_9', 'v_10', 'v_11', 'v_12', 'v_13','v_14', ‘price'

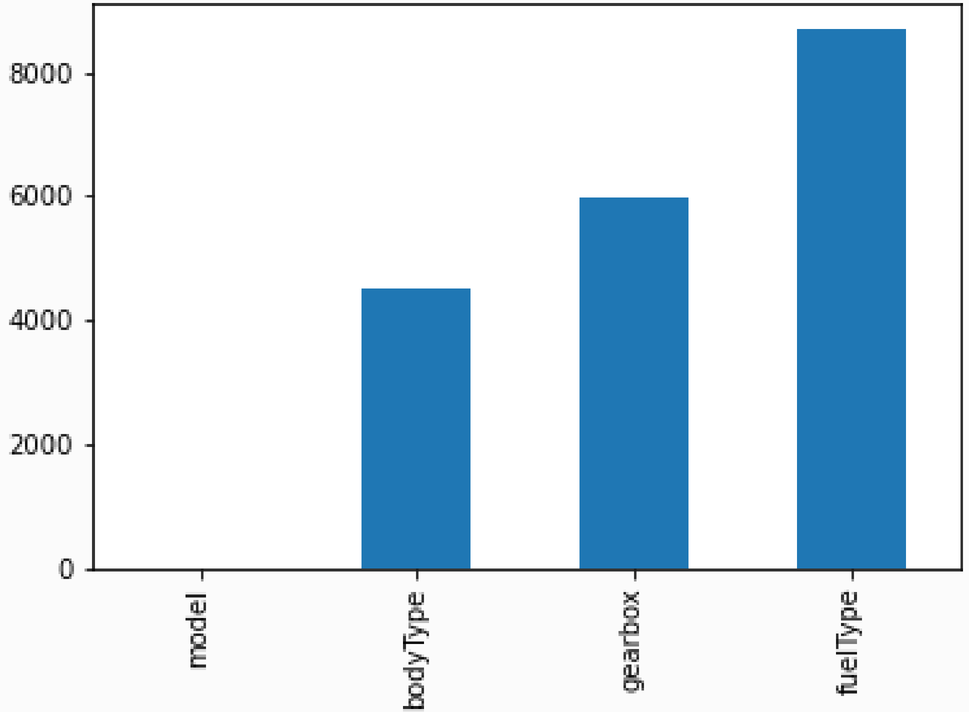
BROWN

I. Project Recap (EDA)

- Pay attention to V0, V3, V8, V12



- Missing value: Using mode 0 to fill the missing value



BROWN

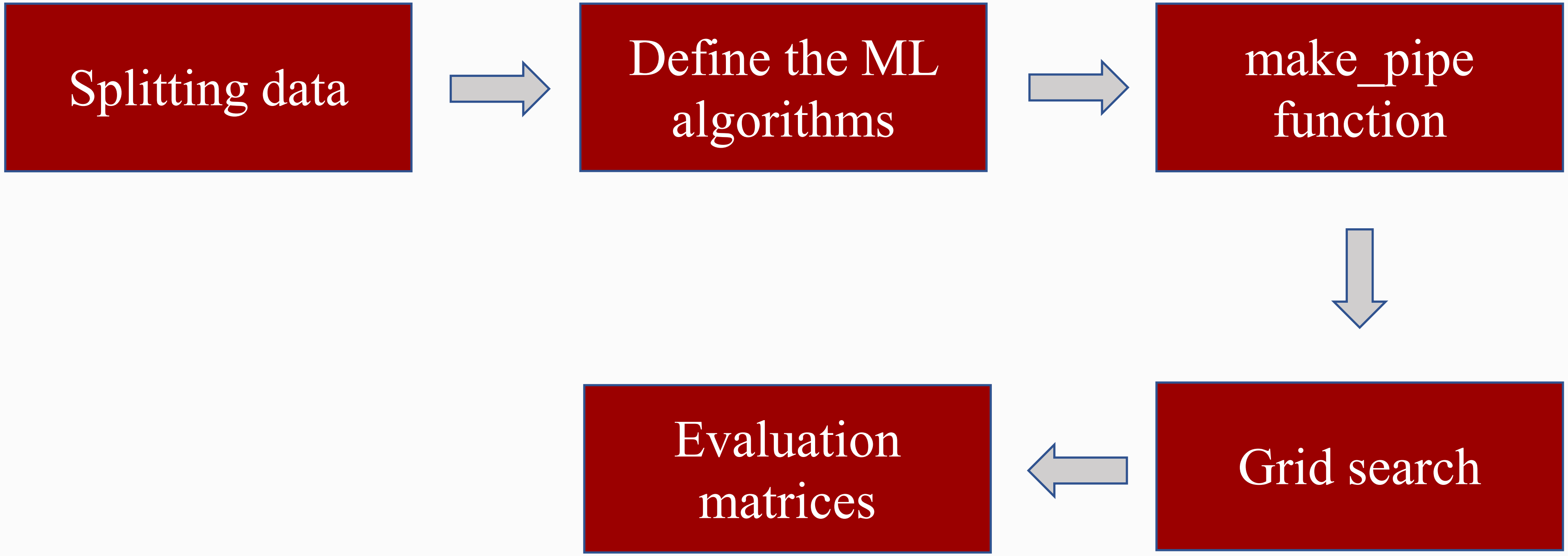
I. Project Recap (Preprocessing)

- Feature Engineering
  - Generate new features
    - Example shows below
  - Preprocessing features with MinMax and OneHot Encoder

```
# using time length: data['creatDate'] - data['regDate'], price decrease if using time increase
data['used_time'] = (pd.to_datetime(data['creatDate'], format='%Y%m%d', errors='coerce') -
                            pd.to_datetime(data['regDate'], format='%Y%m%d', errors='coerce')).dt.days
```

BROWN

II. Cross Validation (CV Pipeline)



BROWN

II. Cross Validation (Algorithms, Parameters)

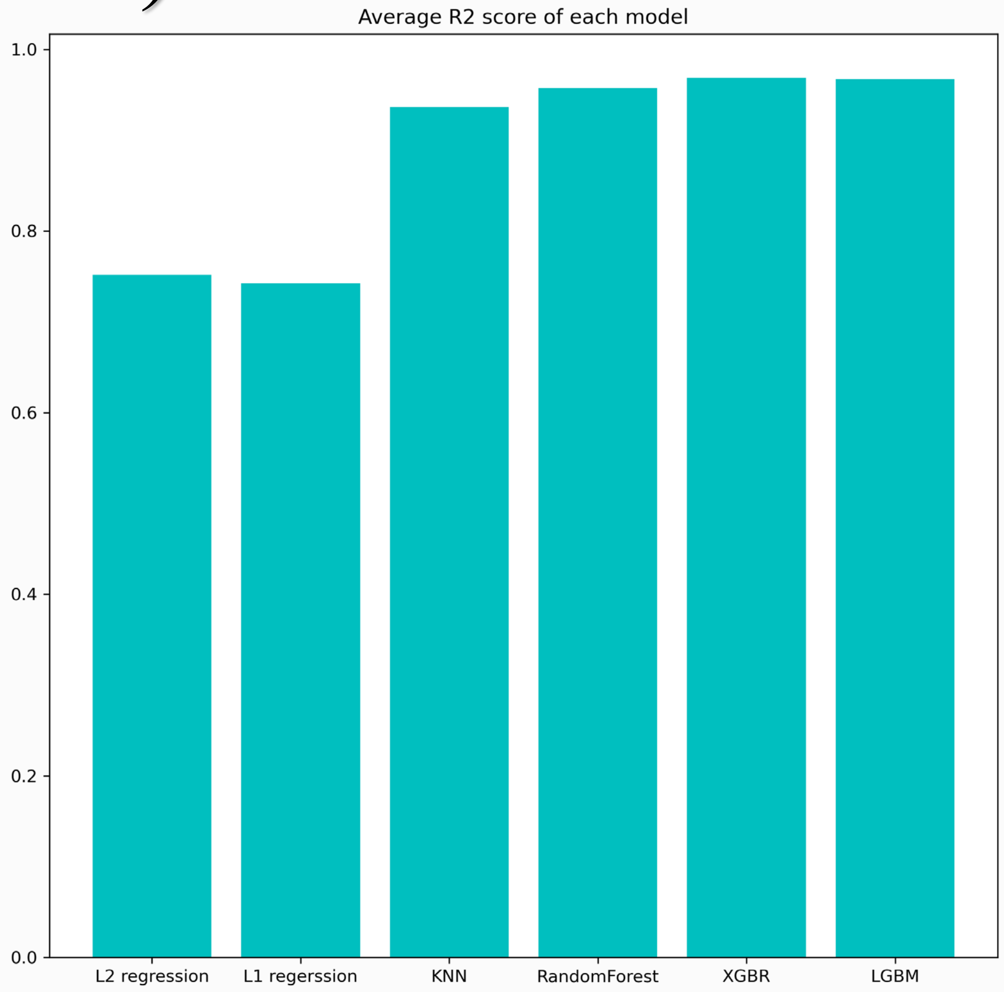
- 6 Regression Algorithms are tried
- Parameters tuned are listed in the table

Model Name	Parameters tuned
Linear Regression with L1 regularization	Alpha: np.logspace(-5,5,25)
Linear Regression with L2 regularization	Alpha: np.logspace(-5,5,25)
K-Neighbor Regressor	n_neighbors:1,11,30,100
LGBM(Light Gradient Boosted Machine) With gbdt (gradient Boosting Decision Tree)	max_depth: -1,1,2
XGBoost	max_depth: 2,3,4,5,8 subsample: 0.75, 0.8
Random Forest	Not tuned

BROWN

III. Results (Model Scores)

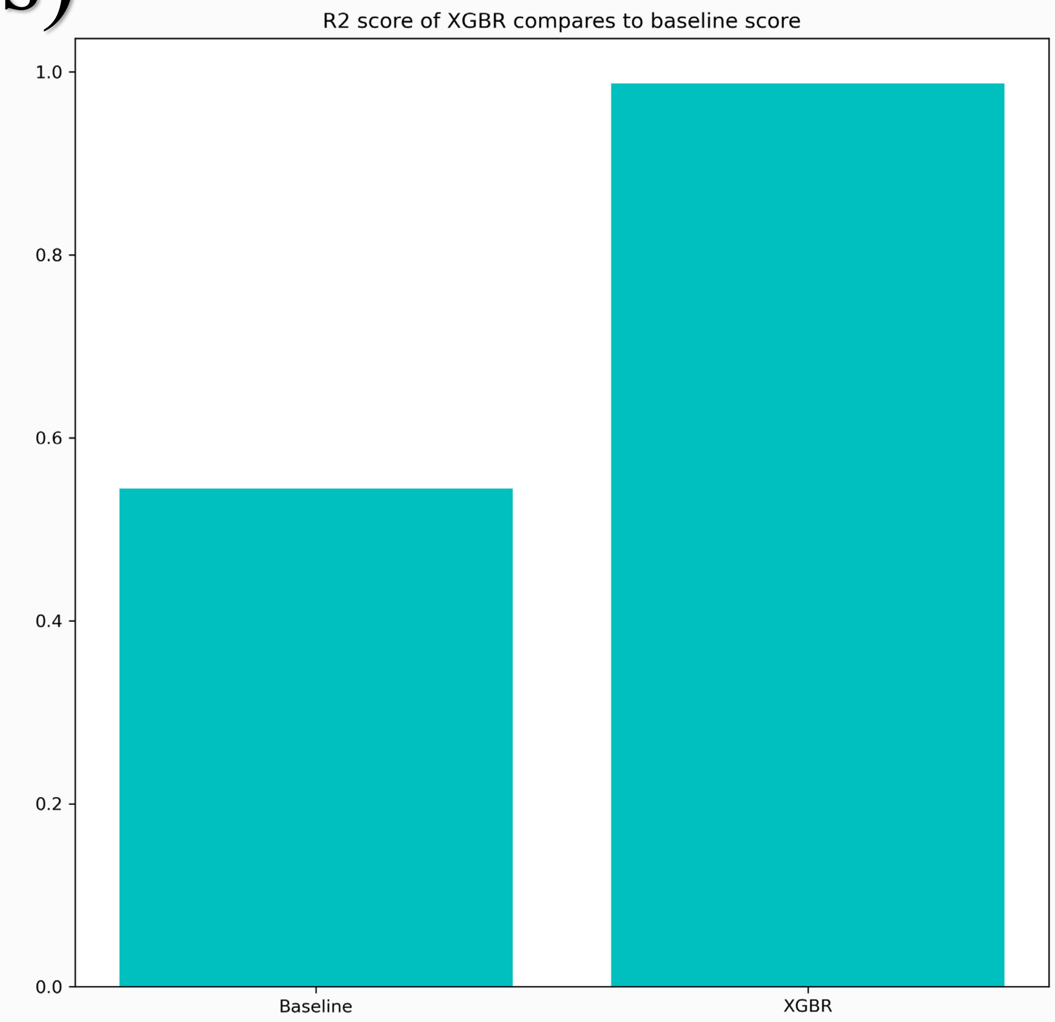
- LR with L1: 0.742
- LR with L2: 0.751
- K-Neighbor: 0.937
- Random Forest: 0.957
- XGBoost: 0.969
- LGBM: 0.967



BROWN

III. Results (Model Scores)

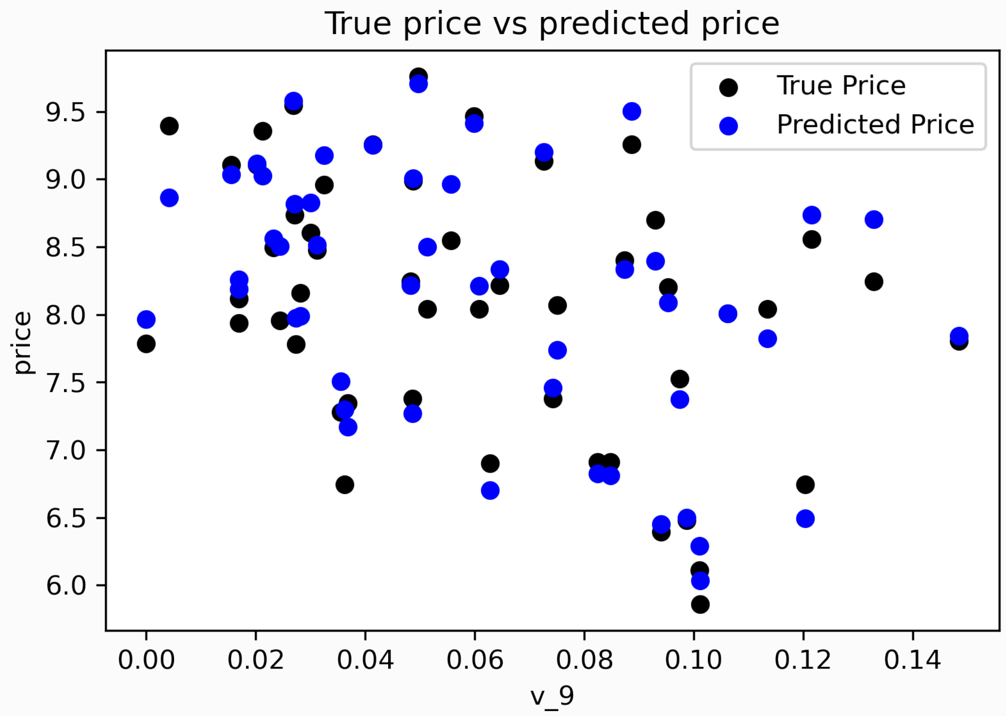
- Baseline: 0.55
- XGBoost: 0.969



BROWN

III. Results (Models Inspection)

- ➢ LR with L1 model: First 50 points. The predicted price are generally away from the true price



BROWN

III. Results (Global Feature Importance)

- Global Feature Importance of Random Forest

Weight	Feature
0.2692 ± 0.0013	new3-0
0.2471 ± 0.0012	new0-3
0.0213 ± 0.0007	new12*year
0.0173 ± 0.0003	new8*year
0.0130 ± 0.0001	new8+3
0.0120 ± 0.0002	notRepairedDamage
0.0108 ± 0.0003	kilometer
0.0102 ± 0.0001	new3+8
0.0079 ± 0.0001	v_14
0.0060 ± 0.0001	new11*year

- Global Feature Importance of XGBR

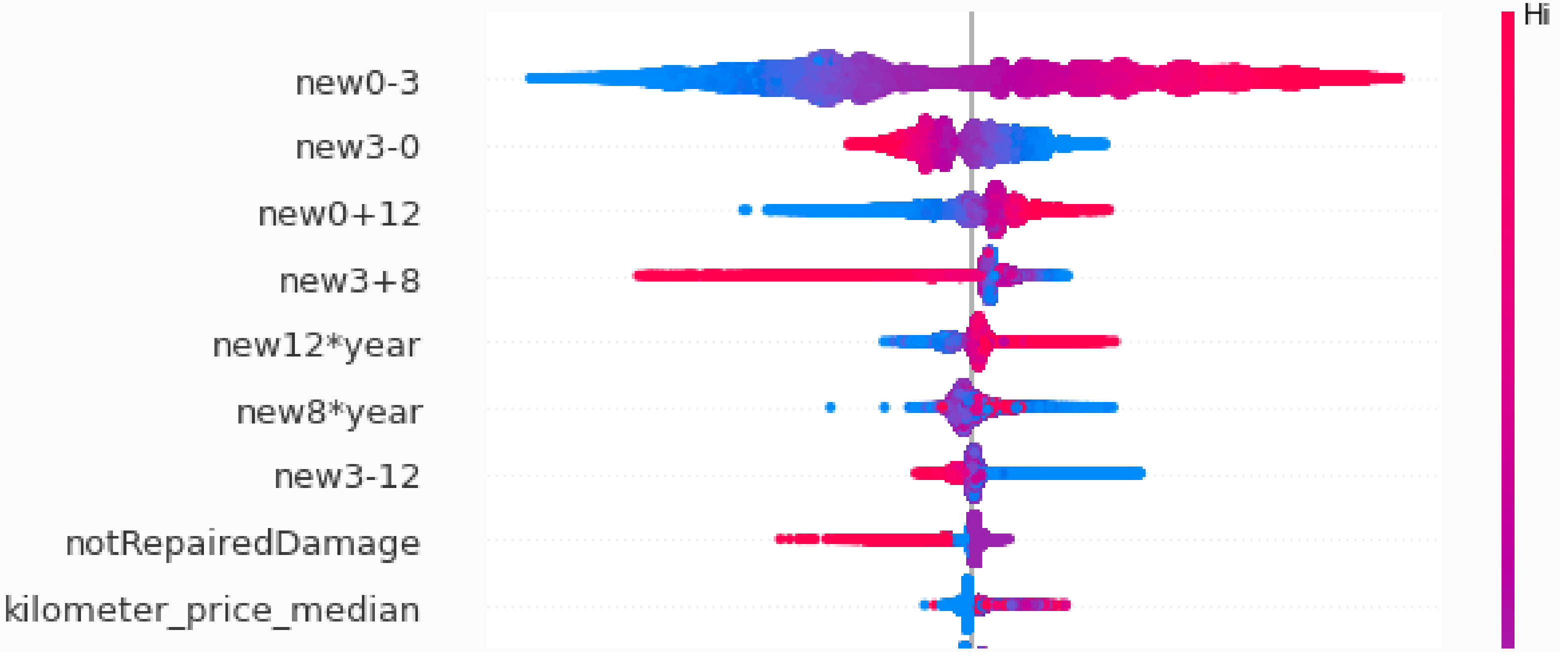
new0*12	0.002930
new12*year	0.005344
new3-12	0.006653
new3-8	0.008650
new12-3	0.008883
new8-3	0.018331
new12-8	0.021881
new0+12	0.032838
new8+3	0.038139
new12+0	0.041154
new3+8	0.206456
new0-3	0.257088
new3-0	0.321593

dtype: float32

BROWN

III. Results (SHAP)

- Part SHAP plot of XGBR



BROWN

IV. Outlook

- Use LGBM rather than XGBoost
- Tunes Alpha in XGBR
- Using model stacking combine XGBR and LGBM
- Buy a 28-cores computer

BROWN

V. Reference

1. “天池_二手车交易价格预测数据分析.” 开发者的网上家园, www.cnblogs.com/cgmcoding/p/13279789.html.

2. 零基础入门数据挖掘 - 二手车交易价格预测赛题与数据-天池大赛-阿里云天池. tianchi.aliyun.com/competition/entrance/231784/information.

3. Lundberg, S. (2020, October 6). *Interpretable machine learning with XGBoost*. Medium. Retrieved December 10, 2021, from https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27.

4. Andrew Lukyanenko. (n.d.). *Predicting molecular properties*. Kaggle. Retrieved December 7, 2021, from https://www.kaggle.com/c/champs-scalar-coupling/discussion/96655

BROWN

Appreciate Your Time
and
Patience

BROWN