House Pricing

Final Presentation

December 2021

Annela Pajumets, Kadi-Liis Kuum,
Kadi Kilgi, Laura Anna Tammesoo

Who we are

- Four master students from Actuarial and Financial Engineering

Kadi

Annela

Kadi-Liis

Laura

Idea of the project

- Kaggle competition
- Model the sale prices
- 79 feature variables to use
  - Training - 1460 examples of different houses
  - Test - 1459
  - USA, Iowa, Ames

The data description and feature engineering

- Decoding NaN values
- New variable - Quarters
- Dummy variables
- PCA

The data sets

- **Regular** - Numerical variables + categorical variables as dummies
- **PCA 1** - PCA on numerical variables
- **PCA 2** - PCA on numerical variables + categorical variables as dummies

Methods

- ❏ Regressions (Linear, Lasso, Ridge)
- ❏ Tree methods (Decision Tree, Random Forest)
- ❏ Ensemble (LR+Lasso+Ridge+DT)
- ❏ Boosting (Adaptive, Bagging, Gradient, XGBoost)

Best results on validation set

- 20% of training set
- RMSE
- **Regular** : ensemble model
- **PCA 1** : XGBoost
- **PCA 2** : XGBoost

Results in kaggle

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(\hat{y}_i + 1) - \log(y_i + 1))^2},$$

- Kaggle uses RMSLE
- Linear Regression **regular** dataset - 4467 place on leaderboard
- Gradient Boosting **regular** dataset - 1786 place on leaderboard

Lessons learnt

- Need to forget the statistics part
- Not used to Python
- Feature engineering was more time-consuming than expected

Thank you for listening!

Results overview

	Linear	Lasso	Ridge	Decision Tree	Random Forest	Ensemble	Adaptive boosting	Bagging	Gradient boosting	XGBoost
Regular RMSE* RMSLE* Kaggle score	31 567 0.1963 0.19424	28 666 0.1224 0.13628	30 826 0.1337 0.14416	46 053 0.1905 0.21339	29 790 0.1246 0.14522	**27 824** 0.1214 0.13625	35 450 0.1727 -	30 999 0.1257 0.14602	28 204 0.1139 **0.13436**	34 794 0.1422 0.13462
PCA 1 RMSE* RMSLE* Kaggle score	40 386 0.197 0.21052	-	41 503 0.177 0.19926	36 829 0.176 0.20865	33 880 0.149 0.18594	-	-	34 468 0.15 0.18618	33 972 0.18 0.18376	**28 414** 0.15 **0.18291**
PCA 2 RMSE* RMSLE* Kaggle score	-	-	32 676 0.14 0.16021	36 218 0.159 0.19796	30 780 0.123 0.16825	-	-	28 452 0.12 0.15224	26 481 0.121 0.15224	**26 376** 0.117 **0.14971**

* Results are from one validation set (20% of training data)

Reference

Picture taken from:

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview

Github:https://github.com/latammes/ML-Project