1861402 Dusica Stepic
1848425 Omirbanu Nurassilova
1852026 Ivana Nastasic

team name: DuBaI
score: 0.11347

Language: Python

Data tidying

- Normalization of the target feature-SalePrice by using the logarithm of the Sale Price values
- Removing outliers where GrLivArea is more than 4000 and log of SalePrice <13
- Removing attributes with a percentage of NAN values greater than 48%
- Resolving multicollinearity by removing feautures that are too correlated
- Removing attributes that contain a high percentage of the same values and don't have a large value of mutual_information(MI)
- Transformation of MSSubClass from numerical to categorical

Feature engineering

- ✔ Adding new features and dropping the features used to derive them:
  TotalSuperficial = TotalBsmtSF + 1stFlrSF + 2ndFlrSF
  TotalBsmtBath = BsmtFullBath + BsmtHalfBath
- ✔ Removing irrelevant attributes, and the ones that are in high correlation with other attributes
- ✔ Removing attributes which have the same values in the majority of rows (more or equal than 90%)
- ✔ Creating dummy variables for the categorical attributes

Modelling and regularization:

1. For model selection k-fold (5) cross validation was used
2. For predictions we chose a ***weighted combination*** of models and ***stacking*** in order to improve accuracy and lower the RMSE:
   - GradientBoosting*0.1+Lasso*0.1+Ridge*0.15+xgboost*0.25+StackingCVRegressor*0.4

   whereby StackingCVRegressor is composed of following models:
   - regressors = (Ridge, Lasso, ElasticNet, GradientBoosting)
   - meta_regressor=ENet
3. Transforming the logarithmed prices back to the original values with exp and storing them as predictions