Predict Duration Time

1/ Overview Data:

This dataset has 398 rows and 7 colums :

Input columns: Product Category, Inspected Samples, Measurement Samples, Workmanship Samples, Number of Item type, Number of Styles.

Output column : Total Time (H)

Get feature from Product Category: There are 85 categories.

Type of this column is string, there are some methods to extract this features such as onehot vector, tfidf or lda. However, in this data , there are only 398 samples so above ways are not fit to apply.

My method used in this project:

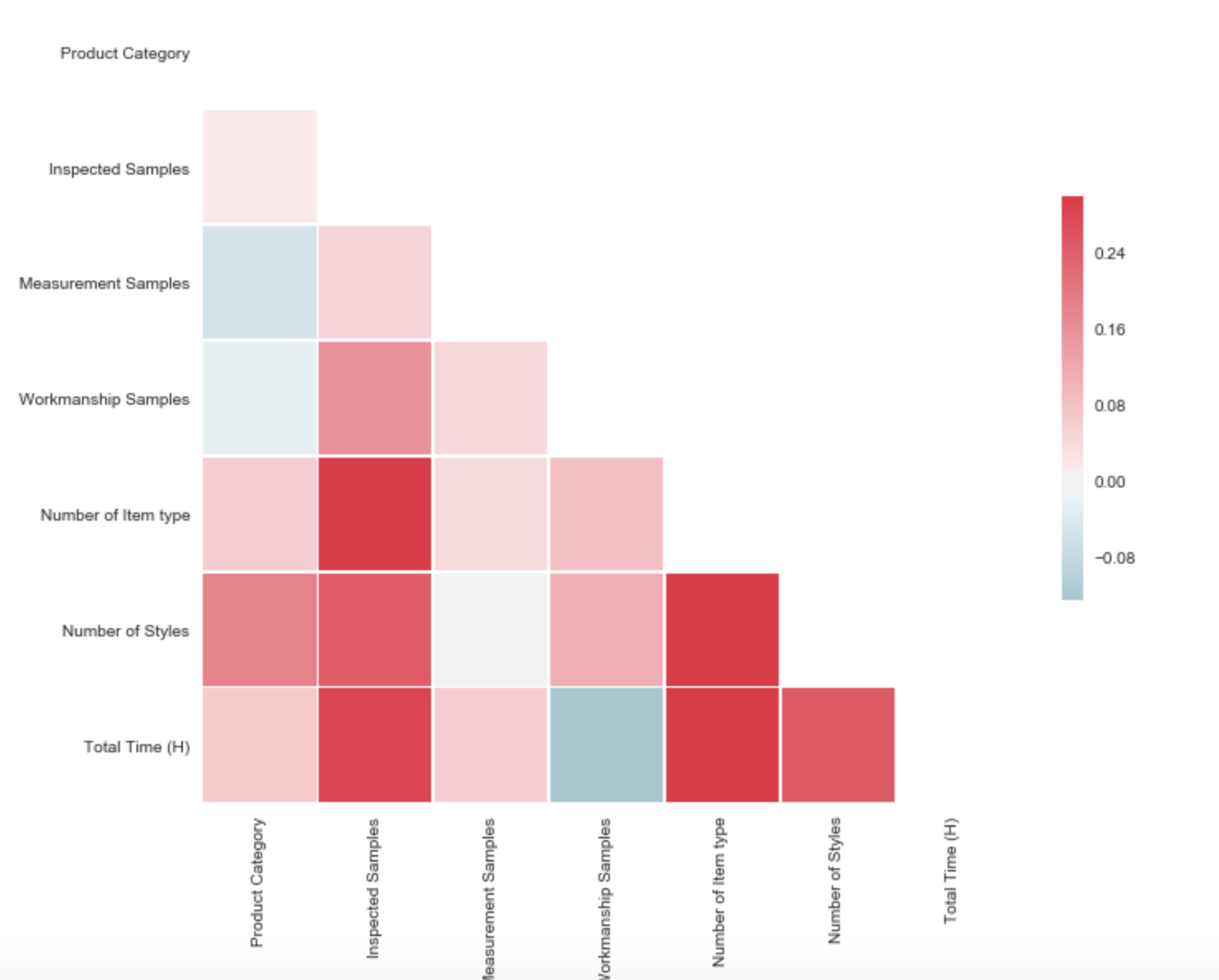
+ Fristly, we have to reduce number of categories by only use the frist component in category. Reduce only 54 categores

+ Second, Build a histogram of category to map each category to p(x = specific category).

Detect Outlier:

Use outfence = 1.5\*(Q3-Q1) + Q3 to detect outlier, robustscaler in sklearn to preprocessing the input data and drop samples which have outlier target.

Correlation of features



H1. Correlation of featues and target

2/ Use 5-fold cross validation to evaluate best type of model:

|  |  |  |
| --- | --- | --- |
|  | RMSE | R2 |
| Elastic | 1.57 | 0.1148 |
| Randomforest | 1.47 | 0.2181 |
| Xgboost | 1.42 | 0.2637 |

H2- The result of t-fold cross validation

Base on this table, I decide use Xgboost to build model.

3/ Result

After tuning I have a xgboost model with params:

XGBRegressor(max\_depth = 3, learning\_rate=0.1,

objective = 'reg:linear', subsample = 1.0,

colsample\_bytree = 0.9, colsample\_bylevel = 0.6,

n\_estimators= 300, reg\_lambda = 101.)

Use early stoping to find number of loop is 164

|  |  |
| --- | --- |
| Features | Measure importance feature |
| **Product Category** | **0.249** |
| Inspected Samples | 0.103 |
| Mesurement Samples | 0.078 |
| Workmanship samples | 0.144 |
| **Number of Item Type** | **0.261** |
| **Number of Styles** | **0.165** |

**Some result of my model on cross validation and training set:**

RMSE\_train = 1.2363968081033632

R2\_train = 0.4677836974721532

RMSE = 1.2230665716247016

R2 = 0.46816380854108774.