Credit Forbearance prediction using XGB and Light GBM models

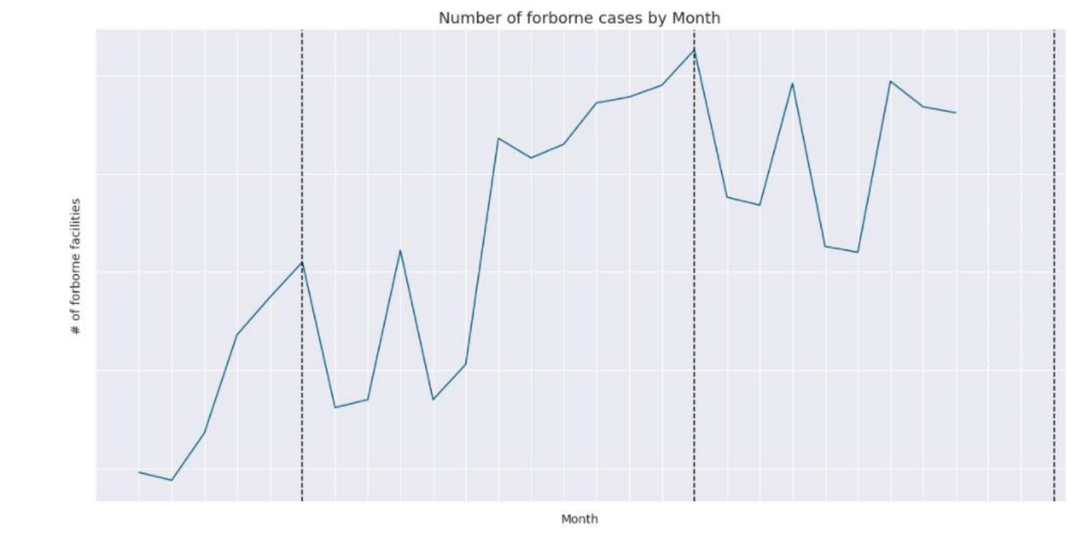
Dalia Breskuvienė
Risk Data Scientist at DanskeBank

Abstract

Bank carefully tracks its weak and vulnerable (W&V) clients segment. One of the ways to become a W&V customer is when a facility becomes forborne. It is essential to follow the "as is" situation and look into the future of W&V segment development. The model aims to predict if the customer's facilities will become forborne during the upcoming six months.

The data cleaning part requires a trade-off between the amount of source data and functionality, as data granularity is on the customer facility level each month when the scope of the data is 24 months.

It was chosen to use the XGB and Light GBM Classification models as they recently showed the best performance on the bank data in other problem-solving tasks. XGBoost and Light GBM are high-performance gradient boosting frameworks based on a decision tree algorithm. Light GBM differs from other decision tree algorithms because it splits the tree leaf-wise instead of the tree depth-wise or level-wise. A complete model would allow risk managers to take strategic actions based on the predictions.



Challenges

Challenge 1

Too big data set for pandas:

- ~10*e6 rows in investigated period
- ~250 initial features

Solution

- To cluster facilities based on:
  - Age/Living area
  - Bank products
  - Exposure

Result

- Number of clusters: 13
- Random sample from each cluster is taken to a new data set

Challenge 2

Imbalanced dataset – there are much less forbearance marked facilities than not forbearance.

Solution

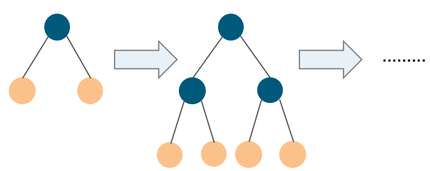
Model parameter scale_pos_weight [default=1] controls the balance of positive and negative weights.

Methods

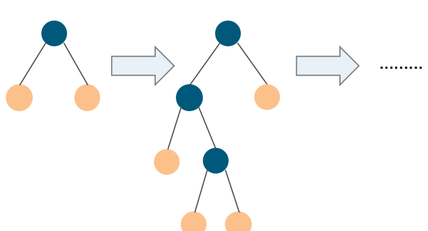
XGB is level wise tree while Light GBM is leaf wise ML model.

The XGBoost is built on a gradient-boosting decision tree algorithm. The model splits up to the defined max_depth hyperparameter and then prunes the tree backward. It removes splits where there is no positive gain.

While XGBoost trees grow horizontally, Light GBM algorithm grows vertically, indicating it grows leaf-wise. Light GBM is not favored for small datasets as it can easily overfit due to its sensitivity.



Level wise tree growth in XGBoost



Leaf wise tree growth in Light GBM

Data Structure

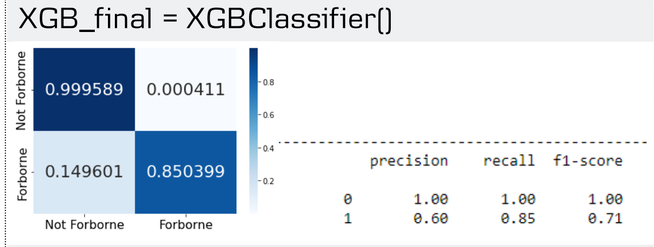
As it is a forecasting problem, the model can not use current financial or risk event information. Data with 6 to 11-month lag and 3 month moving average are used instead.

Risk event related features	• 3 month moving average • 6/7/8/9/10/11 month lag
Financial information	• 3 month moving average • 6/7/8/9/10/11 month lag
External macro features	• 3 month moving average • 6/7/8/9/10/11 month lag
Agreement features	• Initial values

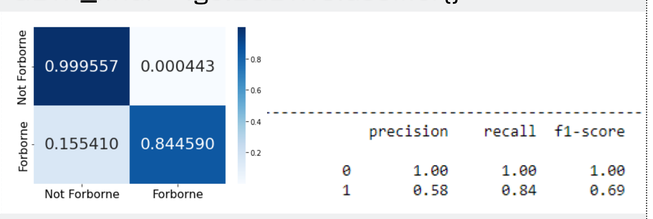
Results

Basic Model vs Tuned Model

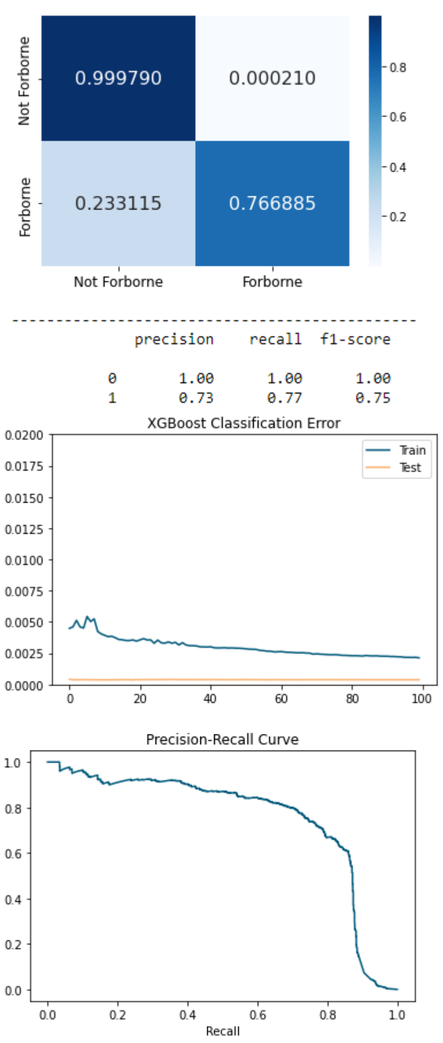
XGB_final = XGBClassifier()



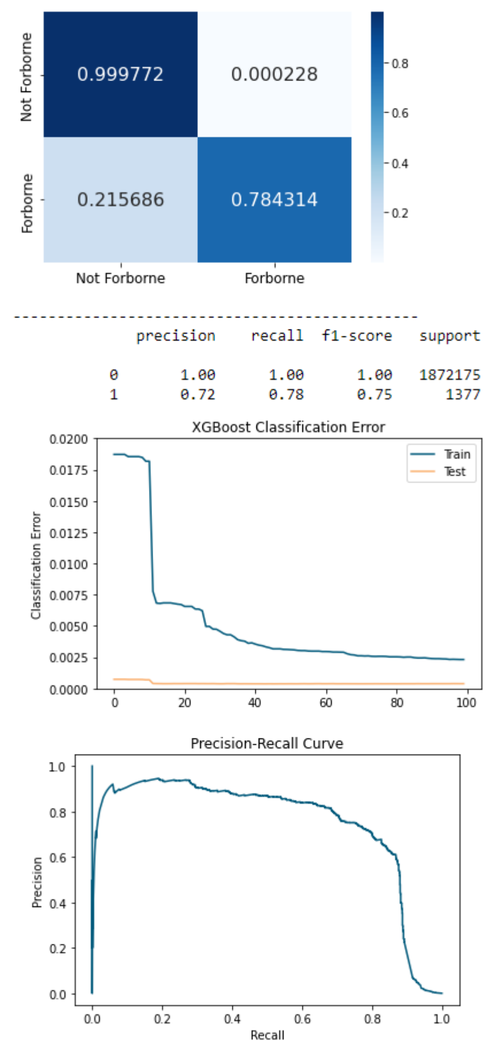
GBM_final = lgb.LGBMClassifier()



XGB_final = XGBClassifier(
objective="binary:logistic"
, scale_pos_weight = 0.05
, learning_rate = 0.1
, max_depth = 4)



GBM_final = lgb.LGBMClassifier(
objective="binary",
scale_pos_weight = 0.05,
learning_rate = 0.05,
max_depth = 4)



XGBoost Model explanation

ELI5 library is used to explain XGB Classifier. It returns an explanation of model prediction as feature weights. Table lists features that contributes to the model. Top 2 features that influence prediction are risk event related.

Weight	Feature
0.1307	Risk_related_f1_rolling_average
0.1285	Risk_related_f2_rolling_average
0.1115	Finantial_f1_rolling_average
0.1049	Risk_related_f3_rolling_average
0.0617	Finantial_f2_lag_6
0.0514	Agreement features_f1
0.0438	Finantial_f3_rolling_average
0.0405	Risk_related_f4_lag_10
0.0199	Finantial_f1_lag_10
0.0179	Agreement features_f2_lag_6
… 102 more …	

https://eli5.readthedocs.io/en/latest/autodocs/xgboost.html

XGBoost Prediction results