Kaggle competition - Home Credit - Credit risk model stability

Description

This competition was one of the biggest competition on the famous Data Science Website Kaggle.com Over 3500 teams joined. More than 5000 participants joined the competition. I have teamed up with 2 data scientist to work on this project.

Objective

The goal of this competition is to predict which clients are more likely to default on their loans. The evaluation will favor solutions that are stable over time. Lack of credit history can lead to loan denial, but data science could improve loan accessibility by predicting repayment abilities. Consumer finance providers use scorecards—statistical and machine learning models—to assess loan risk. These must be updated regularly due to clients' changing behaviors, balancing model stability with performance. Home Credit, since 1997, offers responsible lending to those without credit history, enhancing financial inclusion. Assessing client default risks can help providers approve more loans, aiding those previously excluded. The objective of the "Home Credit - Credit Risk Model Stability" competition on Kaggle is to create a predictive model that can assess the risk of clients defaulting on their loans, with an emphasis on the model's stability over time. The challenge is to develop a solution that remains reliable in the long term, helping consumer finance providers to make more informed decisions and potentially extend credit to those who lack a credit history.

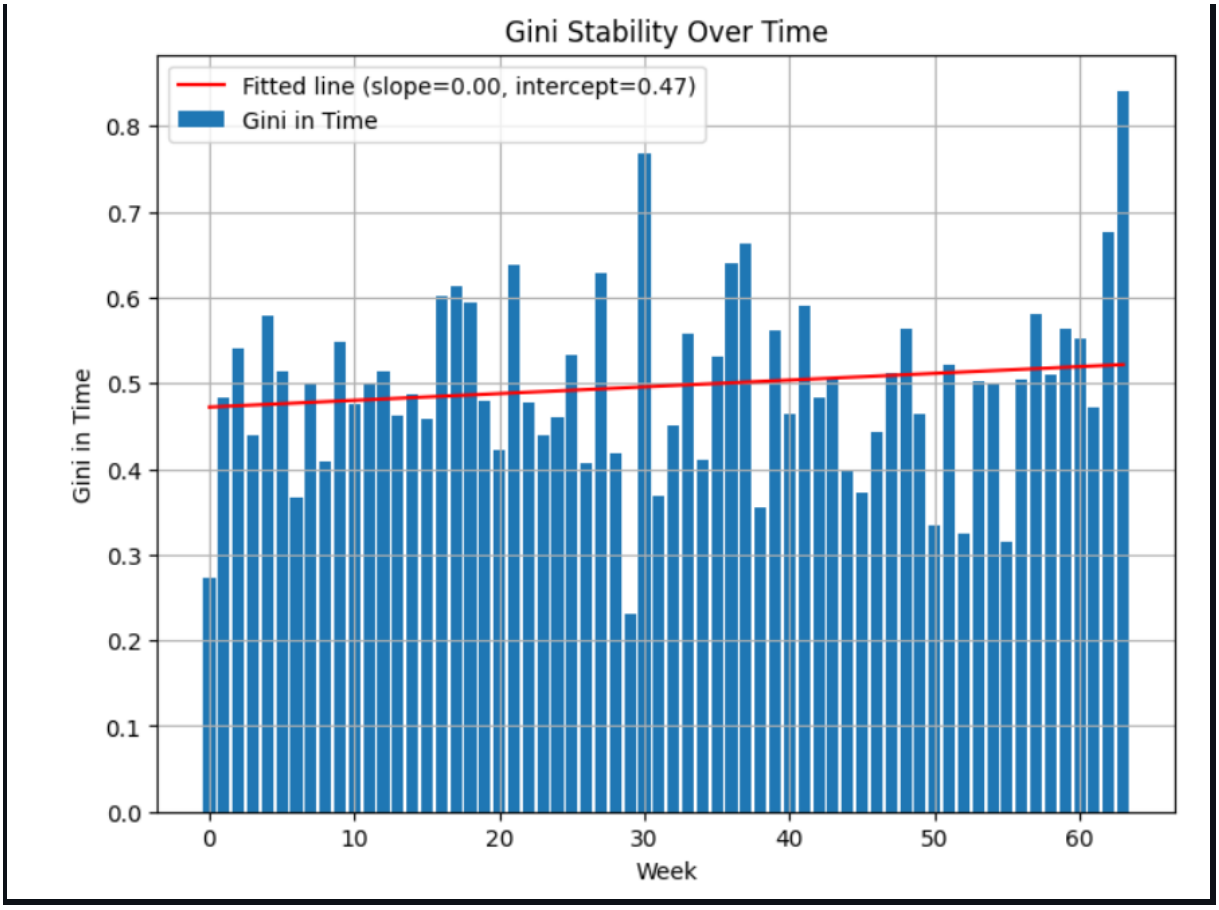
Dataset

A huge training dataset has been provided. Since the dataset comes from real company, the data have contained a lot of discrepancies. The process of data cleaning and preparation has been crucial. Dataset contained data both from internal sources (Home Credit company). Dataset includes several files: information about clients, information about previous loan applications of clients, information from Credit Bureau, information about debit cards of the clients. After aggregation the dataset contained about 1.5 million rows (= loans where the default is to be predicted).

Approach

First, we have experimented with single models. We have tried models like XGBoost and LightGBM. LightGBM showed great results on the CV validation set. We proceeded to tune the model to achieve even better results. We have put some effort into feature engineering, although later we got to know that engineered features does not bring much improvement. Next, we have experimented with other types of single models. We have developed notebook

with RandomForestClassifier, however RF did not show good results. We have created Tensorflow notebook, which turned out to be a great candidate for future ensemble model. Additionally, we have found another high-performance model that was usable for this kind of prediction model – CatBoost. We decided to build an ensemble model consisting of LightGBM, CatBoost and Tensorflow. We have even developed a generic notebook where any ensemble model could be developed, called THESEUS. We were about to develop the ensemble model and tune it properly, when the competition changed the rules.



There have been many disputes by other participants in the competition, whether the evaluation metric is relevant and prone to "metric hacking". "Metric hacking" does not mean any jailbreak. It simply means applying unconventional methods that maximizes the score for particular competition. The "metric hack" itself does not bring any usability in real-life. The competition hosts have changed the rules and allowed for the "metric hacking", before it was prohibited. In this particular competition "metric hacking" was very effective way to improve the score. That's when the effort of the most of the participants shifted from feature engineering into "metric hacking". So did we.

Results

We have placed the 130th in the public leaderboard and the 1069th in the private leaderboard (out of 3858 teams). Our approach was very similar to the approach of the absolute winner of the competition – ensemble model: LGBM, CatBoost, Tensorflow.

Conclusion

Risk assessment is one of the most important use of data science in finance sector. This competition provided real-life data. The fact that we have dealed with real-life data made the project challenging. I believe this competition has provided me valuable experience in the field of loan default risk assessment.

References

https://www.kaggle.com/competitions/home-credit-credit-risk-model-stability