Matthew Walsh

MTH 320 – Chou

Project Proposal

Intro

One project proposal that I would be interested in doing is to try out for myself to work on some finance related projects, with a specific focus on assets risk management and portfolio optimization. Risk management is a tricky thing to work with in finance since most guesses are based on broader macroeconomic factors as well as past credit bureau data to try and figure out if a borrower will default on their loans. The problem with these types of models is that they in general are potentially inaccurate in their predictions due to being based on a long history of financial choices which sometimes fail to capture the true financial position and individual is in at the time. However, with recent developments in the amount of data available to lenders, the opportunity for newer predictions is opening in ways that it did not previously. While estimates previously were based mostly on outstanding debt decisions, new data on a transitionary level such as bank account cash flows, card spending behaviors, rent and utilities payments compared to income, and new gig economy incomes allows us to get a much better view of the entire picture of individual finance. The goal of the research is to try and find whether or not transaction level data helps to improve credit risk models or not, and if any causal relationship is robust enough for it to still be acceptable to use in lending environments with strict regulations.

Methods

Most current credit risk models that are widely used are simple logistic regression models. This is done for many reasons due to their simplicity in understanding, good predictive power, and ability to be implemented easily. For a lender to deny a loan based on merit and avoid potential lawsuits, they must be able to provide an explicit reason. With regression models a lender can point to specific factors in an application that do not fit their lending models and therefore comply easily with regulations. Regression models are also simple to update and keep track of, making them more adaptable to whatever the current market conditions are compared to more complex algorithms. They are also in many cases more than suitable for credit risk, due to the linear types of relationships in credit data. In this case, normal logistic regression can provide an easy and robust comparison to evaluate other types of algorithms. Data transformation will be key for this to be effective. Transactional data such as income variability and overdraft frequency are good at measuring a borrower's short term financial stability but must first be transformed into economically significant variables.

Tree-based machine learning models are something that could be particularly interesting to explore. On the front of credit risk analysis, they can be used as both a means of measuring approvals as well as portfolio monitoring and early default risk prediction. These models can be used to help focus on non-linear relationships that regular regressions might not capture as easily. Using the transactional data from individuals like income and tenure and comparing them with credit bureau and product specific variables, it may be able to find appropriate outcomes like approval cutoffs and risk pricing bands. Beyond that, using instances such as declining cash balances, more requests for cash advances, and payment volatility may help to predict the chances of default.

Tree models could also be used to similar results in the matter of portfolio optimization. One useful factor to view from a non-linear approach is the matter of optimization under constraints. In particular, the minimization of potential loss in the event of low liquidity situations is important. The global economy is in extreme leverage, with high amounts of corporate and private debt. By using non-linear modeling, tree-based models can help to predict potential events where risk may accelerate quickly rather than growing slowly over time. By taking data from previous periods of comparable leverage and stress, it may help to categorize potential events that could trigger losses proportionately larger than the actual percentage loss like what happened int eh 2008 housing crisis.

Packages

R has quite a few useful packages for financial modeling. For general portfolio optimization and risk assessment, there are packages such as portfolio analytics (good for setting constraints, running optimizations, and integrating risk tooling), performance analytics (useful for back test evaluations), and xts / zoo (good for time-indexed data to keep signals aligned). On the front of credit risk analysis there are packages like scorecard, which is good for data partitioning, WOE binning, and is useful in tree-based models as well thanks to its risk feature development monitoring.

On the Front of the tree-based models, there are again many packages which could be useful for modeling and robustness checks. XGBoost is a useful package which can take the inputs of many smaller, less predictive trees and create a more unified model with a singular output. This may be helpful for combining several different variables of transactional data and focusing them into determining credit approval or default risk. Ranger and Random Forest are good for creating fast random forests, good for creating baselines as a means of comparison, as well as robustness checks for non-linear segmentation.

Data

One dataset that could be worth looking into on the matter of credit default risk is called Home Credit Default Risk and is found on Kaggle from user Home Credit Group. The data is comprised of several external data sets containing information about an individual's transactional and behavior statistics, based around a central table of their loan application information. It covers a wide array of different metrics such as loans, installment payments, credit card balances, bureau credit records, and point-of-sale cash balances. These variables can be used in Weight of Evidence or standardized models to create scores in line with regulation expectations in lending practices. This model would be extremely useful in tree-based models, especially using the XGBoost packages discussed earlier, to help find detectable patterns in less specific datagroups and merging them into a single focused explanatory model.

https://www.kaggle.com/competitions/home-credit-default-risk/data

Likewise, the data set "Give Me Some Credit" posted to Kaggle by user Credit Fusion, can be useful for predicting potential credit failure. However, one shortcoming of this dataset in particular is that the data was built around a two-year horizon window (whether or not a person would default within two years, not after two years). The data is also lacking in some of the transactional factors that were discussed earlier as being a potential focus, but it is still useful for many other things. The data is very suitable for logistic regression, covering variables such as debt ratios, age, income, and delinquency counts. Beyond that it can still be useful in tree-based models even without the transactional data in predictors for the effects on defaults after debt ratios hit a certain level or potential links between delinquency rates and income instability.

https://www.kaggle.com/c/GiveMeSomeCredit/data

Potential Research Questions

Do tree-based machine learning models have an advantage over standard logistic regression models in consumer credit risk identification? This would be helpful in identifying potential lenders int he cases where it matters the most with extreme losses. Using boosted tree models as well as logistic regression models you can identify the highest default rates between characteristics as well as the highest loss concentration and use bootstrapping to mark confidence intervals for robustness.

How do liquidity and leverage changes affect nonlinear portfolio losses? This question would be really great to look into, but data sets on this type of thing seemed like they would need to be looked into farther. The data exists in things like regulatory bank filing reports but would likely need to be compiled from multiple files and years to be comprehensive. Using equity returns and liquidity data, a continuous loss event exceeding a certain threshold could be identified to create a tail end analysis of the data leading up to it to try and identify predictors.

Conclusion

Both of these two topics, portfolio optimization and risk management as well as credit risk analysis and default prediction, would make great topics for a capstone project. Looking into aspects of financial analysis previously dominated by regression models this time comparing them with non-linear interactions would provide a new view on delicate topics that could always use extra verification. Using a new modeling approach would compare suability, understanding, as well as predictive power and accuracy. This would look to explain how altering viewpoints on a traditionally one-dimensional assessment method could add new layers of understanding to processes that are done daily and affect people's lives.