MAGENTA

**3001 ML for Finance Initial Modeling Plan**

Hongyi Zheng, Mei Chen, Ken Zeng

**The Business Problem**

Predict the probability that a potential borrower (with assets > €1.5MM) of Banca Massiccia will default on a principal or interest payment for a prospective loan over the next 12 months.

**User Scenario**

This model can help Banca Massiccia to better reduce manual investigation costs, and the bank can use the prediction to further evaluate and offer different pricing and interest rates based on different customers' estimated risk in defaulting the loans. Additionally, the model can help the bank reduce the potential loss if prospective customers have a higher probability of default.

**The Related Data Mining Problem**

This is a supervised classification problem that predicts default probabilities. There are a few ways in investigating this problem: data-driven models, structural models, reduced-form models, etc. We will mainly focus on data-driven approaches and combine some financial instincts to preprocess and select the related features - we can utilize additional information about the market situation as context to our problem. For example, the year 2008 was the financial crisis, and making predictions about that year would be significantly different from a normal year. Therefore we should look to mine basic market metrics, such as inflation rate, unemployment rates, and the performance of the European equivalent of the S&P 500.

How??

**The Unit of Analysis**

The unit of analysis in our project is the default behavior of each company within each fiscal year. The default behaviors of each company in different fiscal years are considered as different records.

**Potential Target Variables**

The output of our final model would be a numeric variable ranging between 0 and 1 indicating the probability of default for a firm in a specific year. Nevertheless, the target variable that will be used to train our classification model would be a binary variable transformed from the "**def_date**" column ("12/31/99"

How will you calculate this ??

indicates that the default date is NA, which means the borrower did not default, other dates indicate that the borrower did default on a specific date).

**Potential Features**

Most of the potential features will be used in our model besides some that will not appear in future data. For example, "**fs_year**" and "**stmt_date**" will not be included in the training and validation set since the model we build aims to predict the probability of default for borrowers in the **future**. Thus, including the fiscal year and statement date as a feature would not improve the model performance. Instead, these features will be used to conduct temporal data splitting of the dataset.

We also notice that categorical features such as "**legal_struct**", "**HQ_city**", and "**legal_struct**" has a large number of potential values, and the number of data points in each category in the demo data set is very limited. However, we can also represent categorical variables as a sample statistic, for example using the CatBoost algorithm, to reduce the sparsity induced if we were to use 1-hot encodings. Whether to include those features or not in our final model depends on the size of the full data set and the distribution of the number of data points within each subclass, which can potentially utilize the correlation matrix to further feature engineer it.

Are there other, simpler things you could try?

Additionally, there are central features that we will pay attention to, such as return on assets '**roa**', profits '**profit**' etc. that are closely related to borrowers' ability to fulfill their obligations. We might need to further normalize the corresponding features to avoid data imbalance in feature weights as well.

How??

**Potentially Model Solutions**

This is a supervised classification problem. If interpretation is of concern, our potential model choices include but are not limited to Logistic Regression Classifier and SVM as well as a simple decision tree. However, in practice, most of the best-performing models often utilize an ensemble of different classifiers to produce more robust models. Examples of these include the widely popular XGBoost, CatBoost, and LightGBM boosting algorithms. These have had a tremendous amount of success in structure prediction problems with heterogeneous features. To further improve the model performance, We can also try out ensembling methods to combine the strength of multiple weakly correlated models to reduce the number of misclassifications further.

How does your proposed approach differ from what you would do for a non-finance problem? How are you using the financial structure of the problem to aid in the formulation??

**How Do the Results Solve the Business Problem?**

By accurately identifying the probability of default for a potential borrower, Banca Massiccane to increase the interest rate and increase the underwriting fee when lending to a high-risk borrower to compensate for the additional risk brought by the higher default risk. If the probability of default is too high, Banca Massiccia could also decide not to lend to this borrower. In this way, Banca Masiccia could effectively

reduce the risk of sustaining heavy losses due to a large amount of default when the economic situation worsens.

How does your proposed approach differ from what you would do for a non-finance problem? How are you using the financial structure of the problem to aid in the formulation??

**Evaluation Metrics**

If we treat this problem as a classification problem, then there is a large amount of class imbalance within the data. ~93 companies that defaulted out of 1000 companies. Therefore if we use a simple naive model that always predicts companies to not default, we will automatically end up with 90% accuracy. Instead, we should use F1 scores which are more robust to class imbalance. Furthermore, we would prefer high recall. Since it's generally safer for a bank to expect a company to default, prepare for it, than to lose all its investments when a company suddenly defaults. In the same vein, AUROC would also be another good metric since it also takes into account how different probability thresholds perform.

What do the other metrics provide that is not availoasble in the ROC an alysis?

We care about the model's performance on unseen data in the future, therefore we should structure our training and validation set to evaluate this. For example, for the training set, we can only include companies that defaulted before 2010, validation with companies between 2010 and 2011, and test set using companies that defaulted between 2011-2012. This would allow us to evaluate how well the model can generalize into the unseen future. In addition to this, since the company statements were taken at different times, we should also evaluate the model's performances on different subsets of the data, e.g group by region, the legal structure of the firm, etc to check for potential hidden biases within the model. Ideally, the model should have a similar performance across all subsets and any deviations from this would be a sign that we need to investigate the data more closely.

**Potential Extension**

Nowadays the interpretability of machine learning models is becoming increasingly important. Instead of merely getting a number representing the probability of default, Banca Massiccia may also want to look into the model to see why our machine learning model makes such decisions. Therefore, after constructing a well-performing model, we will apply the LIME interpreter on top of our model so that Banca Massiccia could investigate whether the reasons behind a decision made by our model match the human prior.

Is there an easier way?

Preprocessing?
Variable selection?
Pervious work by other researchers?

ECONOMIC INTUITION / INTERPRETATION??