Project Proposal (Part I)

Noah Keogh (njk6) S01264950

The goal of my project is to build a model that will be able to predict if someone will experience financial distress in the next two years. This model will use a variety of financial metrics in order to determine a given person's likelihood. One possible business application of this model is that it could be used by banks and other financial institutions to assess how risky someone would be to loan to. For example, if someone is likely to experience financial distress in the next two years then it would be unwise to give credit to them as it may be unlikely that they will be able to pay it back in the future. If, however, someone were to be unlikely to experience financial distress in the next couple of years then they may be considered a safer investment as compared to someone with a high risk of experiencing financial distress. Not only could this model be used by banks in determining whether someone is a risky investment, but it could also be a feature in a credit scoring model. In this case, people with a lower risk of financial distress would have a higher credit score as compared to those who have a high risk of financial distress. Thus, we can see that there are a range of business applications. In terms of this project, the goal will be to determine how likely someone will be to experiencing financial distress in the next couple of years. The applications to building a credit scoring model will be out of the scope of this project as that would require other financial metrics about an individual that I do not posses. An additional goal of the project will be to see if there are groups of people that have different financial tendencies or behaviors. For example, do people with a large number of loans have high incomes. The goal will be to see if there are any behavioral tendencies and whether these behaviors can be used to distinguish people who are likely to experience financial distress from those who are unlikely to.

In order to build a model such as the one described above, I will be using a publicly available dataset from Kaggle [1]. This dataset contains 11 features related to an individual's finances. One of these features can be considered a label. This feature is binary in nature and indicates whether a given individual has experienced financial distress, which is defined as having a payment that was over 90 days delinquent. Thus, since the dataset contains a feature that contains a label, I will be able to consider both supervised and unsupervised learning models to achieve the business objective. In addition, since there are over 10 features, there may be some features that may be more useful than others in terms of predicting risk of future financial distress. Thus, I could perform dimensionality reduction in order to determine which features or dimensions are most important. The dataset contains around 150,000 samples in the training data, so there is are a large number of samples to learn from.

If I were to implement this strategy, and use it to asses the risk of financial distress for an individual, I would need to make some assumptions about the data. I would need to ensure that the individuals in the training data were all sampled from the same population and that this population is representative of the population that the model will be used on. In addition to this assumption, I must assume that each of the samples are independent and identically distributed so that all individuals are independent from one another and that they are randomly sampled so there is no bias introduced in the sampling of the dataset. If I do not make this assumption, I would not be able to make a successful model as the biases introduced during the sampling phase would be carried through to my model. I must also assume that the numbers presented are accurate in nature. I must assume that these are not self-reported numbers, but were rather assessed by an organization and that these assessments of an individual's finances are consistent. If I do not make this assumption, it would be hard to build a model as the data would be inaccurate, and any model trained on it may also contain the same inaccuracies. Since I had no part in the collection or sampling of this dataset, these are the assumptions I must make.

Credit scoring or financial riskiness models have existed for a very long time, and, thus, have been used for many years without the use of machine learning models. In the 18th-century in the United States, storekeepers would get secured loans based on their character which was verified by people (such as neighbors) who would vouch for them [2]. In the 19-th century these models had not progressed much beyond, and were still biased and not based on many financial metrics [2]. It was not until the 20th-century that consumer credit reports were created and financial loans and credit worthiness were based more on financial metrics. In the 1970s, a large change came about when credit scores were released to the public for the first time [2]. Thus, it is clear that financial riskiness models have existed for a long time, and, therefore, have existed without the use of machine learning models.

In today's modern world, however, the major credit score bureaus (whose scores are used by banks and institutions for determining the financial risk and credit worthiness of an individual) incorporate machine learning algorithms in their models. The algorithms and models used are proprietary to each credit bureau so we cannot know exactly which ones they are using, but credit bureaus such as FICO have stated that they have been incorporating machine learning into the calculations of their scores for over 30 years [3]. Though the use of machine learning is prevalent in credit scoring models, an individual does not need to have access to the scores or machine learning models in order to determine if they are susceptible or likely to experience financial distress. Simply analyzing ones finances may reveal how likely one is to experience a serious delinquency. Though machine learning models are not required for these simple analyses, machine learning models could aid individuals in identifying trends or patterns in their finances that could indicate that they are likely to experience financial distress, long before it is clear to an individual. Thus, some problems involving financial distress could be solved without the use of machine learning; however, there is large benefit in using machine learning models to identify patterns of financial distress, as these models can identify and see patterns that an individual cannot. Nowadays, all of the major credit bureaus appear to

be using machine learning algorithms in their credit scoring models and, have been doing so for some time. Thus, this project aims to mimic these models, and identify how relationships between different financial metrics can be used to asses the likelihood of financial distress.

References

[1] Show me some credit. https://www.kaggle.com/competitions/GiveMeSomeCredit/overview.

[2] Sean Trainor. Your credit score's long history, from espionage to algorithms. https://time.com/3961676/history-credit-scores/, Jul 2015.

[3] Machine learning and fico scores. https://www.fico.com/en/latest-thinking/white-paper/machine-learning-and-fico-scores.