**EEE485 PROJECT PROPOSAL**

**Introduction**

For our term project, we have decided to utilize a set of supervised and unsupervised learning algorithms to implement a company bankruptcy prediction system. In modern times, companies have a large set of defining financial parameters which are open for public use. These parameters provide crucial information on the financial situation of a currently operational company. When accounting firms and banks are evaluating such a company for various reasons, such as for loan applications, they require a solid understanding of the company’s financial standing. For such operations, a system that can evaluate a company’s standing and predict whether a company is at risk of going bankrupt or not with a high reliability would be advantageous. The motivation here is to implement such a system by data analysis methods and machine learning algorithms.

**Methodology**

For development, we will be using the Python programming language. Within Python, we will utilize, at minimum, the following four non-machine learning libraries:

* Matplotlib: Basic statistical data visualization.
* Seaborn: Improved and easier to use data visualization.
* Numpy: Mathematical tools such as matrices and vectors.
* Pandas: Reading CSV files.

We are planning on using a dataset with high amounts of features per company. For this reason, we will perform dimensionality reduction with some unsupervised learning methods such as “PCA” and “k-Means Clustering”. Moreover, considering how the amount of companies that go bankrupt per year is much smaller than the total amount of companies that are operational, it is likely that the dataset(s) we use will have a heavy imbalance towards companies that did not go bankrupt. This creates a classic “imbalanced classification” problem. To combat this, we are planning on using oversampling to generate a more balanced data. Then, the balanced dataset will be used to train one of the classification algorithms given below:

* Logistic Regression
* Gradient Boosting
* *k*-Nearest Neighbors

**Datasets to Be Used**

We plan on utilizing a dataset called “Taiwanese Bankruptcy Prediction Data Set [1]” which we found on the “UCI Machine Learning Repository”. The dataset contains 96 features for 6819 companies listed in the *Taiwan Economic Journal* [2]between the years of 1999 and 2009. From our initial observations, we see that the data is heavily imbalanced as we foresaw. Therefore, we will probably end up using an oversampling method. Additionally, we will have to be careful while examining the data since the period includes the 2008 financial crisis which might have imposed financial troubles on some companies that were otherwise stable and not at risk of bankruptcy.

**Expected Challenges**

At this point, the challenges we expect are:

* Efficiently combining the supervised and unsupervised learning algorithms we mentioned earlier.
* Fixing the imbalance of the data without damaging the characteristics of the dataset.
* Lowering the runtime of the execution.

**Conclusion**

To sum up the proposed project, we intend on building a classifier that will aid banks and accounting firm get a better grasp of the future security of a company by predicting whether this company will go bankrupt or not. To do this, we will implement a combination of supervised and unsupervised techniques.

**References**

[1] (2020). Taiwanese Bankruptcy Prediction Data Set | UCI, [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction>. [Accessed: Feb. 15, 2020].

[2] (2020). Taiwan Economic Journal, [Online]. Available: <https://www.finasia.biz/>. [Accessed: Feb. 15, 2020].