CPS 853 Project Report
Stage2 Progress
Company Bankruptcy Prediction

Andy Liang
500833782
Computer Science
Ryerson University
Toronto, Ontario

Tusaif Azmat
500660278
Computer Science
Ryerson University
Toronto, Ontario

I. Motivation (Data Discovery)

The project provides us with the classification problem; it is the suitable choice for Big Data analytical problems. So, we choose Company Bankruptcy Prediction. The purpose of the bankruptcy prediction is to assess the financial condition of a company and its future perspectives within the context of long term operation on the market. Choosing this project will help us to apply the techniques learned and it will help us better understand all the concepts in the Big-Data Systems course.

A. Problem

The objective for this project is to test big data to build machine learning models to classify or identify companies that will file for bankruptcy as per their financial situations. We want to predict if the company will go bankrupt based on the input attributes and identify it as bankruptcy yes or no.

B. Significance and Challenge

The dataset is about bankruptcy prediction of Polish companies. The data was collected from Emerging Markets Information Service, which is a database containing information on emerging markets around the world. The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013. The challenge is the data is highly imbalanced when it comes to prediction class.

C. Prior Work

We choose this project from the dataset provided on Kaggle [1] website. There is some work done but we will use this dataset and apply the machine learning algorithms to further enhance the accuracy of our predictions.

II. Method (Data Prep)

We used the Azure Databricks analytical framework which is cloud-based data engineering tool used for processing and transforming massive quantities of data and exploring the data through machine learning models.

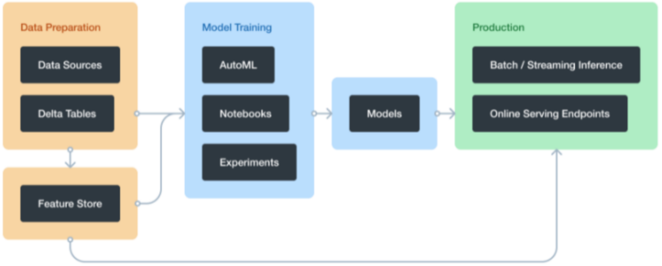

Figure 1: Databricks Machine Learning Model

We used all the techniques which were available to us along with those we learned during the course of big data systems. We applied supervised learning techniques to both the classification and the regression algorithms. There are some algorithms in particular we used are as follows:

1. Naives Bayes classifier
2. Linear Regression.
3. Logistic Regression.
4. Decision Tree.
5. Random Forest.
6. K-Nearest neighbors (KNN)
7. Support Vector machine.

III. Intended Experiments (Model Planning and Building)

We started our project with data collection and for that we use the dataset provided to us through Kaggle [1] dataset. After downloading the dataset we perform some data analytical processes such as preprocessing, analysis and evaluation based on our experiments. We perform the following methods:

A. Dataset discussion

The dataset contains eight years of companies financial records that were obtained by the financial institution. Financial institutions used that data to keep track of companies that declare bankruptcy based on the information collected. Furthermore banks design their financial server based on such information about the companies. The dataset is highly unbalanced with a low percentage of bankrupt companies within several records of normal company’s records. The positive class (Bankrupt) accounts for 4.48% (1126 bankrupt out of 25121 records used for training dataset) of all records.

B. *Pre-Processing and data Normalization*

The dataset used here was already split for us in terms of what is to be used for training and testing purposes but we tried further splitting the training dataset to further train test sets. When splitting the datasets, we first mixed the samples randomly. Then we split the data into training, validation, and test files by 70%, 15%, and 15% respectively.

B. Feature Extraction

We have 66 possible columns to use in our predictions. There are lots of missing values in many column attributes. Feature 'Class' is the target variable with value 1 in case of Bankrupt and 0 otherwise.

C. Specific Learning Algorithms

For the Bankruptcy Prediction Classifier, we have tested multiple different learning algorithms to see which would perform better. Initially we tested classical learning algorithms like Naive Bayes and Logistic Regression and later tried some sophisticated algorithms that gave us promising results.

IV. Results and Milestones

We were able to classify the companies which will declare Bankruptcy with relative accuracy with the models we employed. On both datasets, all models were able to classify with greater than 85% accuracy, with top models, RNN performing greater than 90% plus accuracy with some models on datasets. Furthermore, we were able to determine the companies most likely will declare bankruptcy and our model will predict for new datasets as well.

A. Milestone 1: Data cleanup (Data Preparations) (By Early April)

Feature extraction will be done along with some other matrices.

B. Milestone 2: Training pipeline and evaluation (end of first week April)

As the data set is highly imbalanced, it was a challenge to work with it.

C. Milestone 3: Train and evaluate models (after first week April)

For this task we used different dataset to test and evaluate prediction models.

C. Milestone 4: Failure Analysis and Model Refinement (By mid April)

By

C. *Milestone 5: Final Report with Presentation (April 19, 2022)*

By this data we will have our final report ready with our final model predictions. The code and implementation for this project will be available for evaluations.

V. References

[1] “Company Bankruptcy Prediction.” https://www.kaggle.com/competitions/company-bankruptcy-prediction/data.