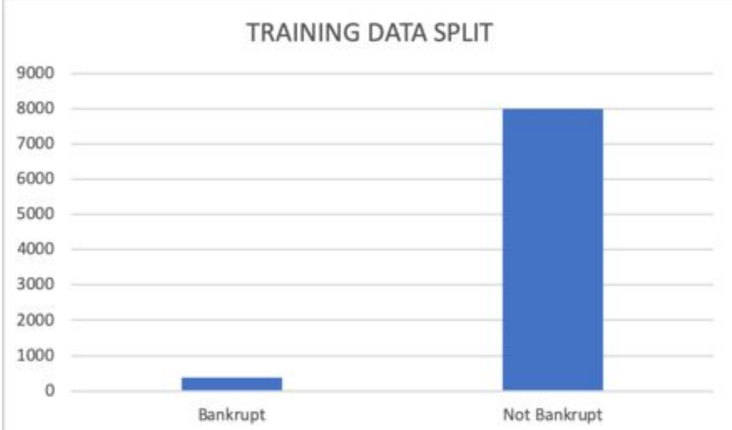
CORPORATE BANKRUPTCY PREDICTION

PREDICTING

Economic crisis of 2007 highlighted the need for a market sustainability. We need predictive tools to better predict such events to prevent such catastrophic occurrence from happening again. The significance of predicting financial distress is to develop a reliable predictive model that will provide a means to measure and predict the financial condition of a corporate entity. The goal of the project is to identify the best classification model in terms of accuracy and performance for predicting the bankruptcy of corporations using various statistical forecasting techniques like Logistic Regression, SVM, Neural Network and Bernoulli Naive Bayes which we all learned in the class along with Ensemble Boosted Trees.

DATASET

To evaluate the performance and accuracy of various statistical forecasting techniques we are relying on raw dataset[2] which is hosted on UCI Machine Learning Repository. The dataset is about bankruptcy prediction of Polish companies in manufacturing sector. The motivation in choosing this repository was because since 2004 Poland saw many manufacturing sector going bankrupt. The research sample consists of bankrupt and functionally operating companies as of 2013. Final Scrubbed dataset had 2,101 rows(corporate entities) 65 columns(financial health indicators).



The data that is available was classified in 5 cases depending on forecasting period. we considered evaluating on third year of forecasting data to maximize the number of training examples, percentage of positive and negative instances in them as well.

FEATURES

In our dataset, we have 64 features indication the financial health of the corporate entity and one label column to indicate bankruptcy status after three years. Due to the incomplete nature of data, it required a generous amount of effort to massaging and scrubbing to be useful for analysis. We followed the 80/20 hold off rule for training and testing our models. In training data, we have total 8,402 instances(firms), out of which 384(4.5%) represents bankrupted companies, 8,018 (95.5%) firms that did not bankrupt in the forecasting period. In test data, we have 2,101 instances(firms), 111(5.2%) represents bankrupted companies, 1,990(94.7%) firms that did not bankrupt in the forecasting period.

MODELS

Since the goal of the the project is to identify the best classification model for predicting the bankruptcy of corporations, we took following models for our classifications. Each model needed its fine tuning for parameters to achieve better accuracy. To overcome incomplete features, we filled those features with mean value of those respective features.

Models	Train Score	Test Score
LogisticRegression	0.952392288	0.946692
SVM	1.0	0.947168
NeuralNetwork	0.954296596	0.947168
BernoulliNB	0.765412997	0.755831
XGBClassifier	1.0	0.526416
LGBMClassifier	1.0	0.70871

Table 1: Accuracy Score for models

We have used l2 regularization in all of the models. There were model specific fine tuning that we had to do namely neurons and layers in neural network, choosing kernels coefficients in SVM, choosing Bernoulli Naive Bayes etc. Few of the default parameters were not giving us optimum results

RESULTS

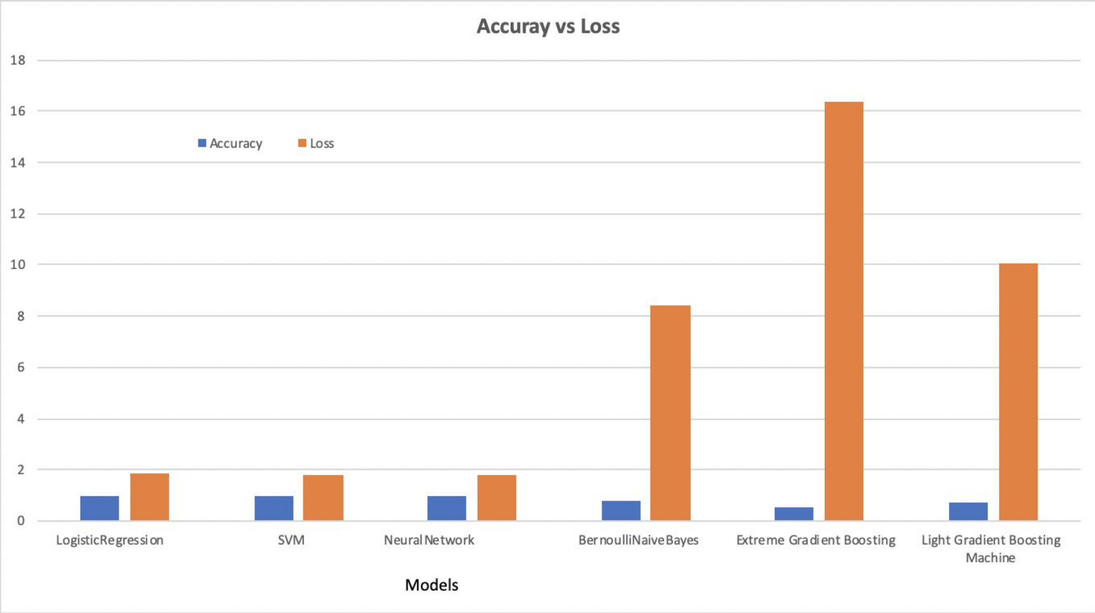


Figure 1: Accuracy vs Loss for all Models

Our observation from our project was that logistic regression, SVM and Neural Network performs way better than any other model in terms of better accuracy and less loss. Surprisingly, we observed that ensemble gradient boosting has higher loss and low accuracy relative to other models.

DISCUSSION

We tracked log loss in all our models. Since it made more sense with our problem statements.

$$-\log P(y_t|y_p) = -(y_t log(y_p) + (1 - y_t) log(1 - y_p))$$

- SVM is a clear winner with 94.7% accuracy and log loss of 1.82.
- Followed by Neural Network with near same accuracy and loss.
- Followed by Logistic Regression with near 94.6%accuracy and log loss of 1.84.
- Followed by Light Gradient Boosting Machine with near 70.8%accuracy and log loss of 10.06.
- Followed by Extreme Gradient Boosting with near 52.6%accuracy and log loss of 16.35.

FUTURE RESEARCH

We were very surprised with bad performance of ensemble boosted methods, which clearly showed high variance and over-fitting in training. Therefore, Future steps should start with replication of latest research[1] which involves enabling ensemble boosted methods with synthetic feature generation. Post replication, next step should be to find ways to elevate the performance even further.