Class Action Lawsuits Prediction Models

Team 8

Xiongfei Ding, Xinlu Tu, Man Zhu, Yuhe Zhu

Introduction

In our model, three steps are taken to predict the security class actions frequency and the damage amount. For the frequency part, we calculate the filed frequency and the settled frequency based on three models: logistic regression, linear discriminant analysis, and random forest. Models are selected based on the AUC score. In filed frequency, models are selected for each industry (see Appendix A for the details); in settled frequency, since all these three models generate similar results, we use the average of the probabilities generated by all these models to increase the stability and accuracy. For the damage amount models, we use MAPE as the criteria to choose among linear regression, PCA regression and neural network model. Since linear regression and PCA regression generate close MAPE values, we decided to use the predicted average by these two models. We create an RStudio App to implement our models, which can predict the SCA frequency and the damage amount of a portfolio of companies.

Data Cleaning

We did not include security class action data for 2000 in our models because we decided to use prior year to predict current year and there are no financials for 1999. Then our group used "id" and "year" to merge two data sets, Historical Class Action Data and US Public Company Financials and found out that there are around 600 observations in Historical Class Action Data, which are not included in US Public Company Financials data set. In order to achieve a complete dataset, our group used S&P Capital IQ to find out the missing financials and add them in the Financials data set. (The extra data can be found in TEAM8/Extra Data.xlsx.) In addition, we found four industry classification categories in Capital IQ: "Industry", "Sector", "Primary industry" and "SIC industry". In order to unify the classification of the industries, we chose "Industry" with least number of classifications, which will result more filed class actions in each industry. Thus, the result will be more accurate. Besides, we delete the "Thrifts and Mortgage Finance" classification as well. Finally, we used SAS to create three cleaned data sets, filed.csv, settled.csv, and damage.csv for building models. Our group use SAS to clean the data.

Modeling

Our model contains two parts, the frequency part and the severity part. Therefore, the expected class action loss can be predicted by getting the product of the estimated frequency and estimated conditional severity. All the relevant codes can be found in TEAM8/T8_damage.R, T8_filed.R, and T8_settled.R.

- **Frequency model**
  This contains two models which predicts the filed probability and settled probability respectively. We used discriminant analysis, Random Forest and logistic as our potential

models, then we used ROC curve as our selection criteria when choosing the proper model.

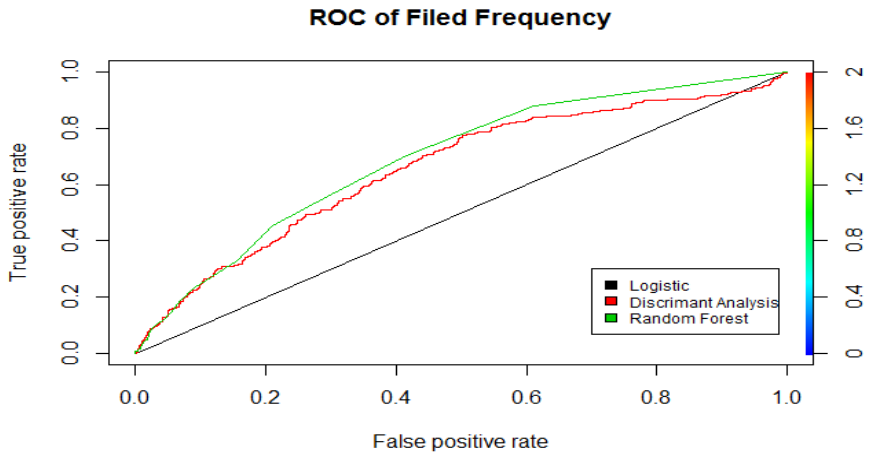
- **Filed Frequency Model**
  First we classified the data into industries, and then we applied the three potential models and the selection criteria individually to each industry. This enhance the model accuracy by introducing various models to different industries. However, for a few industries, the data are too small to be built a model on. Therefore, we use overall model to predict the results. Thus, in this model, we took industry model as our prior choice with overall model as a substitute.

- **Settled Frequency Model**
  The procedures are basicly the same as building the filed model. The only difference is that because the settled data are not significant enough to be divided into industries, therefore it is just a overall model. The dataset we used for this model are companies that has already been filed, which indicates it is conditional probability.

- **Severity Model**
  We first divided the dataset into training dataset and testing dataset. To build models, we used linear regression with box-cox transformation and variable selection, principal component analysis (PCA), and neural network as modeling methods for the training dataset. Then, based on the testing dataset, we calculated mean absolute percent error (MAPE) as selection criteria to decide our prediction models. We might improve our models by dealing with the negative prediction values instead of assigning zero.

Results

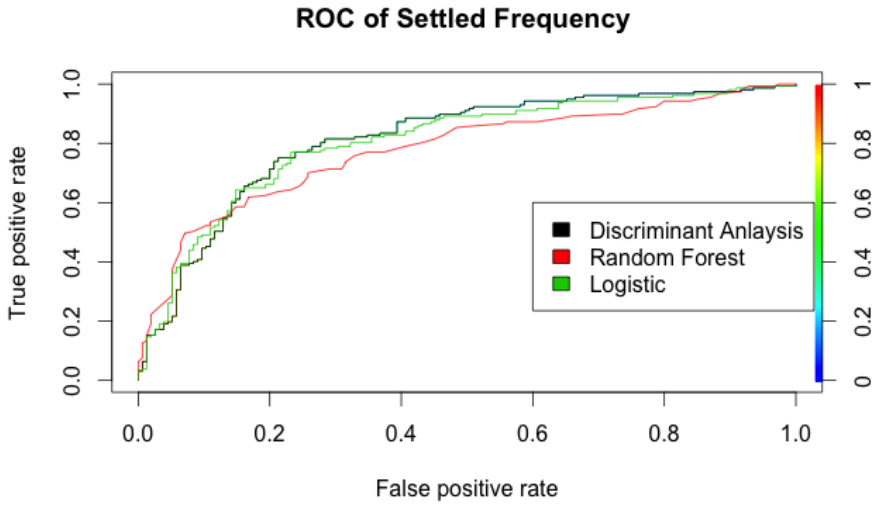
- **Filed Frequency Model**
  The graph below shows the ROC curve of the overall model. As you can see, the results of Linear Discriminant Anlaysis and Random Forest are close, which are much better than that of Logistic Model. Thus, we decide to use the average of the probabilities generated by these two models for those industries which do not have more detailed models.



Next step, we perform all these three models on the rest industries, and select the model with the highest AUC score. Please see Appendix A for the specific information about the models for each industry.

Most of the estimated probabilities are less than 5%. This is reasonable because the odds are low in the provided data sets. It is worth to mention that the ROC curve of a perfect model would be a vertical line at False Positive Rate=0.0. Most of our models do have good ROC results, but others, like the graph above, still have some space to improve.

- **Settled Frequency Model**
  The following ROC curve indicates that the AUC score of the three models we used generate similar results, which are so close that we decided to use the average of the estimated results from the three models as our final probability. Overall, most of the frequency is close to 1, which means once a class action is filed, it is very likely to be settled.



- **Severity Model**
  The linear regression model and principal component analysis model both have relatively low MAPE value (Mean Absolute Percent Errors), whereas the neural network model generates high MAPE value. We decided to adopt both the linear regression model and principal component analysis model, and the average prediction value of the two models is our final prediction.

Model	MAPE
Linear Regression	3.860277457
Principal Component Analysis	3.463520362
Neural Network	14.64839728

Implementation

Please follow the instructions step by step:

1. Open RStudio (it cannot be run in R environment) and install the following five packages:
   randomForest, ROCR, pROC, neuralnet, MASS, shiny
2. Extract AXIS.zip and TEAM8.zip (in Windows system, a new folder will be created to hold the original folder), and put the extracted folders (the original folder not the new folder in Windows system) in your RStudio work directory (it is different from setting the folders as your work directory)
3. Type “library(shiny)” in your console
4. Type “runApp(“AXIS”)” in your console (runApp is similar to read.csv and AXIS folder is considered as a whole)
5. After the App is open, click “Choose File” and input the data set of the companies you want to predict. (We have a sample test file AXIS/data/test_1.csv. The test file must have the same variables as those of our sample test file.)
6. Then you will get all the results from the app.

If any problem occurs, you can check the implementation codes in AXIS/server.R or email us. We are more than happy to help.

Appendix A

Industry	Model
Aerospace and Defense	lda
Air Freight and Logistics	lda
Airlines	lda
Auto Components	lda
Automobiles	lda
Banks	Overall
Beverages	lda
Biotechnology	randomForest.formula
Building Products	randomForest.formula
Capital Markets	Overall
Chemicals	randomForest.formula
Commercial Services and Supplies	randomForest.formula
Communications Equipment	randomForest.formula
Construction and Engineering	lda
Construction Materials	Overall
Containers and Packaging	lda
Distributors	Overall
Diversified Consumer Services	lda
Diversified Financial Services	Overall
Diversified Telecommunication Services	randomForest.formula
Electric Utilities	randomForest.formula
Electrical Equipment	randomForest.formula
Electronic Equipment, Instruments and Components	lda
Energy Equipment and Services	glm
Food and Staples Retailing	lda
Food Products	lda
Gas Utilities	lda
Health Care Technology	lda
Healthcare Equipment and Supplies	randomForest.formula
Healthcare Providers and Services	randomForest.formula
Hotels, Restaurants and Leisure	randomForest.formula
Household Durables	lda
Household Products	lda
Independent Power and Renewable Electricity Producers	Overall
Industrial Conglomerates	Overall
Internet and Catalog Retail	lda
Internet Software and Services	randomForest.formula
IT Services	lda
Leisure Products	randomForest.formula
Life Sciences Tools and Services	glm

Machinery	glm
Marine	glm
Media	lda
Metals and Mining	randomForest.formula
Multi-Utilities	glm
Multiline Retail	glm
Oil, Gas and Consumable Fuels	lda
Paper and Forest Products	lda
Personal Products	randomForest.formula
Pharmaceuticals	lda
Professional Services	lda
Real Estate Investment Trusts (REITs)	Overall
Real Estate Management and Development	lda
Road and Rail	lda
Semiconductors and Semiconductor Equipment	randomForest.formula
Software	randomForest.formula
Specialty Retail	lda
Technology Hardware, Storage and Peripherals	randomForest.formula
Textiles, Apparel and Luxury Goods	randomForest.formula
Tobacco	Overall
Trading Companies and Distributors	randomForest.formula
Transportation Infrastructure	Overall
Water Utilities	randomForest.formula
Wireless Telecommunication Services	randomForest.formula