Financial Predictive Analysis

Report prepared for WalletHub

By

Chun-Wei Lo

September 29, 2017

Executive Summary

The objective of this report is to make predictions on a financial classification problem and find significant factors contributing to the accuracy rate of models. We conducted PCA for dimension reduction and utilized feature selections approach respectively before implementing machine learning models. Our main goal is to compare the model performance based on these two different settings. Classification algorithms used in such analysis included Support Vector Machine (SVM), Random Forest, Logistics Regression and Gradient Boosting Decision Tree with parameters tuning.

1. Introduction

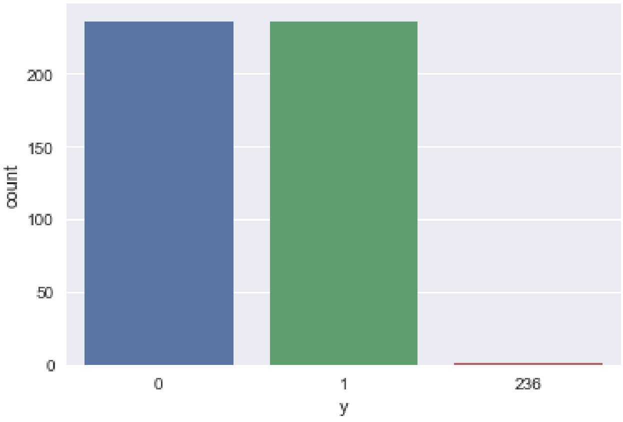
1.1. Dataset Description

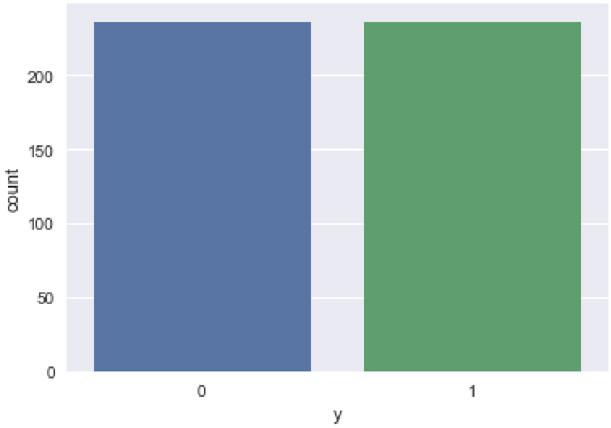
The sample totally includes 444 variables and 473 observations. Our dependent variable is binary variable “y”. It is obvious that we have high-dimensional dataset for predictive analysis.

1.2. Exploratory Data Analysis (EDA)

In statistic, exploratory data analysis (EDA) is an approach to analyze date set and summarize their main characteristics with visualization methods. Generally, we would use correlation matrix to see the relationship between the predictors and response variables or histogram graphs to check the distributions of features.

However, since we were not able to know the representation for each feature, we plotted a bar chart for checking the distribution of target variable. Based on the figure below, we can find an outlier existing in our original sample. After removing it, we can see such dataset have equal number of instance in each class and there is no need to employ sampling techniques before modeling since it was balanced dataset already. If it was highly biased data, including sampling methods is needed to make it more balanced.





2. Methodologies

Our analysis mainly consists of two parts. The first part is performing PCA and the other part is employing Recursive Feature Elimination with cross-validation as our feature selection methods. We further included widely used classification machine learning algorithms, including Support Vector Machine (SVM), Random Forest, Logistics Regression and Gradient Boosting Decision Trees, based on these two settings individually and compared the performance of different models.

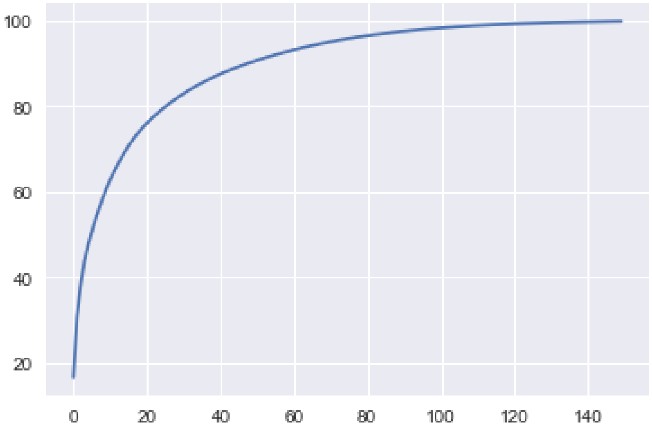
We firstly divided the original sets to a train (80%) and test (20%) split of our dataset for evaluating the performance of an algorithm on the problem. Train dataset is used for fitting a model and test dataset is used for evaluating the final models. In addition, we also made 10-fold cross validation on parameter tuning for SVM, Random Forest and Logistics Regression. After getting the best combination of estimators within the models, we applied them to test dataset for final comparison.

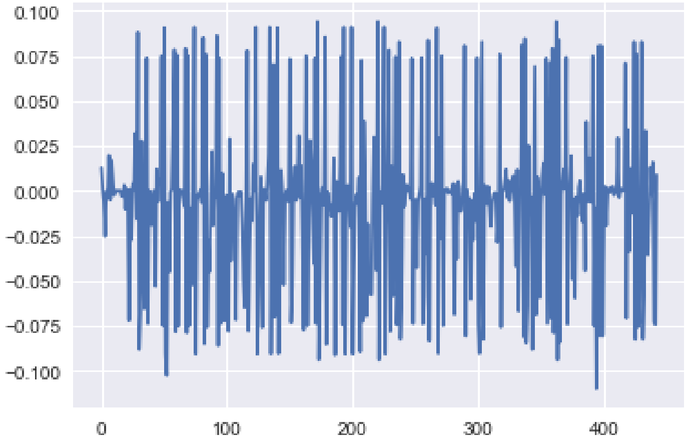
2.1. Principle Component Analysis with Machine learning models

2.1.1. PCA

Principle Component Analysis (PCA) is a dimensionality-reduction technique that is often used to transform a high dimension dataset into a smaller-dimensional subspace prior to running a machine learning algorithm on the data. Each component is a normalized linear combination of the original predictors.

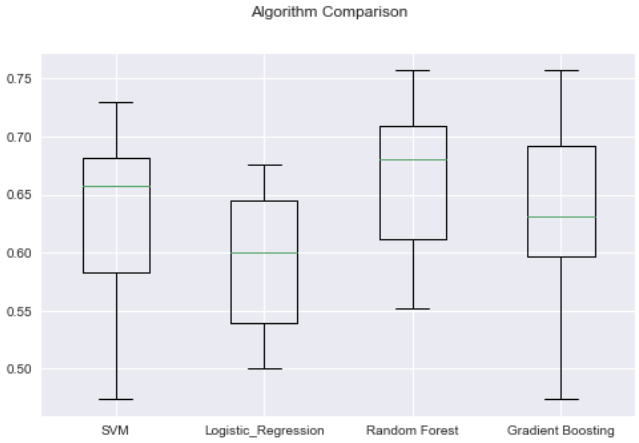
Based on the left figure, we determined to pick 48 as the number of components since those 48 components can explain over 90% of the overall variability after checking the explained variance table. We tried to know what represent the first component from the figure on the right. However, it's hard to interpret the tendency of such component for now. We would use the data after transformation from PCA for algorithms developing.





2.1.2 Result

We used GridSearch on parameter searching with 10-fold cross validation for SVM, Logistic Regression and Random Forest. According to the boxplot shown as below, it is obvious that Random Forest performed better compared with other three models, followed by SVM. However, the variance of SVM is high as well.



We further applied those models to test dataset for measuring the final performance of machine learning models. We can conclude if we conducted PCA beforehand, Random Forest with parameter tuning performed the best in this case, with highest model accuracy rate and roc score followed by Logistics Regression with parameter tuning.

Classification Models	Accuracy	ROC Score
Random Forest with parameter tuning	0.646	0.653
Logistics Regression with parameter tuning	0.645	0.642
Logistic Regression with parameter tuning	0.623	0.632
Gradient Boosting	0.622	0.621

2.2. Feature Selection with Machine Learning models

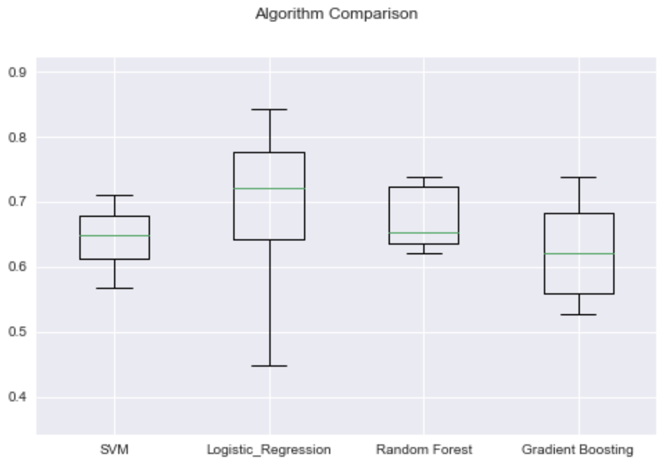
2.2.1. Recursive Feature Elimination

Machine learning works on a simple rule-if we put garbage in, we will only get garbage to come out. By garbage here, we refer to noise in data. This becomes more important when the number of features are very large. In our case, we have totally over 400 features, it is critical to drop some not unnecessary variables for creating an algorithm.

Recursive Feature elimination (RFE) is an optimization algorithm which aims to find the best performing feature subset. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination. We selected the top 48 features into the models. You can find those key features in attached file.

2.2.2 Result

We used GridSearch for parameter searching with 10-fold cross validation for SVM, Logistic Regression and Random Forest. According to the boxplot shown as below, it is obvious that Logistics Regression did the best job among these four algorithms, followed by Random Forest. However, the variance of Logistics Regression is the highest as well.



We further fitted those models on test dataset to evaluate the final performance of machine learning models. We can conclude that if we did features selection beforehand, Logistics Regression with parameter tuning performed the best in this case, with highest model accuracy rate and auc score followed by SVM with parameter tuning.

Classification Models	Accuracy	AUC
Logistics Regression with parameter tuning	0.719	0.705
SVM with parameter tuning	0.688	0.674
Gradient Boosting	0.624	0.611
Random Forest with parameter tuning	0.624	0.611

3. Conclusion

Based on the comparison table, it is apparent that Logistics Regression with parameter tuning performed the best among multiple models. We can conclude that feature selection is useful techniques for improving the performance of the model. Moreover, if we can get involved experts who have domain knowledge in financial field, it would be more helpful to extract relevant predictors contributing to response variable.

	Classification Models	Accuracy	AUC
PCA	Random Forest with parameter tuning	0.646	0.653
	Logistics Regression with parameter tuning	0.645	0.642
	Logistic Regression with parameter tuning	0.623	0.632
	Gradient Boosting	0.622	0.621
Features Selection	Logistics Regression with parameter tuning	0.719	0.705
	SVM with parameter tuning	0.688	0.674
	Gradient Boosting	0.624	0.611
	Random Forest with parameter tuning	0.624	0.611

4. Questions

4.1. Your best regression equation

Our best model is Logistics Regression with optimized parameters as below

```
LogisticRegression(C=0.5, class_weight=None, dual=False, fit_intercept=True,
intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
penalty='l2', random_state=12345, solver='liblinear', tol=0.0001,
verbose=0, warm_start=False)
```

4.2. An appropriate measure of model performance

For classification problem, there are a myriad of metrics that can be used to evaluate predictions, including classification accuracy, area under ROC curve, confusion matrix and report. In our case, we picked classification accuracy and AUC score as our evaluation metrics since there are an equal number of observations in each class. However, we should be careful using accuracy rate since in real world balanced data is rarely the case.

4.3. Why is it important to balance your dataset?

If we don't balance our dataset beforehand, classification algorithms would go for the majority rule since receive most of information from majority class data. For instance, if the ratio of "0" class to "1" class instance is 80:20 instead of 50:50, we may have accuracy paradox problem. Even though we have excellent accuracy, such as 90%, it is possible that the accuracy is only reflecting the underlying class distribution. In other world, we can only predict those "0" class accurately.

There are some methods to deal with imbalanced training data, such as changing our performance metrics (precision rate or recall rate, it depends on the problems) or resampling dataset (over-sampling minority class data or under-sampling majority class data).

4.4. What method did you use to select the most predictive independent variables?

We used Recursive feature elimination for selecting predictive features. Given an estimator that assigns weighs to features, such as the coefficient of a linear model, RFE is to select features by recursively considering smaller and smaller sets of features. Building a model using the remaining attributes and calculates model accuracy. RFE is able to work out the combination of attributes that contribute to the prediction on the target variable (or class).

4.5. Did you divide the data into build and test sets? Why or why not?

We divided our data into train and test sets for evaluating the performance of an algorithm purpose. The training dataset is used to prepare a model, to train it. We pretend the test data is new data where output values are held out from the model. After we gather predictions from trained model on the input from the test dataset and compare them to the withheld output values of the test set. Generally, we need to exclude the information from test set as training models.

In addition to train and test splitting, we should use k-fold cross validation for parameter tuning purpose. With repeated k-fold cross-validation, it will help us get a handle on how much variance there is in different algorithms.

4.6. Why do you believe that your selected measure of model performance is most appropriate?

As we mentioned above, since we have equal number in each class, it is appropriate to use classification accuracy as a measure of model performance. In addition, we also include Area under ROC Curve (AUC), which represents a model's ability to discriminate between positive and negative classes, in our analysis. An area of 1 represents a model that make all predictions perfectly, while an area of 0.5 represent a model as good as random. In our case, the results of both evaluation metrics are consistent, which gave us more confidents in the performance of final model.

4.7. Given more time, what suggestions do you have to potentially make your model more predictive?

- Understanding of each features and domain knowledge

  In order to make our models more predictive, it would be helpful to incorporate with people who have intimate domain knowledge and have experiences in this finance. They can give us some advices on feature selection, interpretability of model and any other relevant factors.

- Parameter tuning for RFE model

  We didn't tune the parameter of RFE currently due to the time limited. However, if we can loop the parameter in RFE for getting the optimized number of features selected, the models should be improved as well.

- Hire me

  With a business background, I am a data enthusiast who aims to uncover hidden insight and drive decision making for application in the IT, e-commerce, Financial, and Fashion industry by using quantitative skills and analytical mindsets.

  I thrive in rapidly growing, innovative organizations. I am passionate about solving business issues and analyzing problems with the help of statistical modeling and a machine learning approaches.

5. Reference:

[1] http://scikit-learn.org/stable/modules/feature_selection.html
[2] https://medium.com/@aneesha/recursive-feature-elimination-with-scikit-learn-3a2cbdf23fb7
[3] Khalid, C., et al. (2013). Classification Techniques in Blood Donors Sector–A Survey. EProceeding of Software Engineering Postgraduates Workshop (SEPoW) 2013, Universiti Teknikal Malaysia Melaka
[4] https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/