K-Nearest Neighbors (KNN) versus Naive Bayes(NB) on predicting bankruptcy of a company

Tharun Guduguntla | Machine Learning

1. Motivation for the analysis

My goal for this model comparison is to check the performance of both the models on this dataset and determine the best performing model in predicting bankruptcy. Taking the research paper [1] into consideration for comparing the model performances with similar parameters.

2. Dataset and Initial findings

- Taiwanese Bankruptcy Prediction data set has been collected from the UCI Machine Learning Repository [4] which contains the bankruptcy information of the companies from the year 1999 to 2009. Data consists of 95 variables and 1 target class label with 6819 observations. Our binary class label indicates 1 as the company went bankrupt and 0 when the company is stable. We don't have any missing values in the dataset. Most of the features were numerical with only 2 categorical variables.
- Data is normalized using the normalization technique and this data will be used for building the models.
- Our dataset has huge imbalances with only 220 companies that are bankrupt(1) and 6599 companies which are stable(0) which can be seen in the bar chart Figure 1.
- To make this a balanced dataset, SMOTE technique has been used which made our binary class labels balanced by inserting the synthetic data into the original imbalance dataset.
- While we have huge variables which affect the companies' fate, we are only considering the top 8 correlated variables in our analysis to make the model more explainable. You may check the correlation values of the features shown in Figure 2.

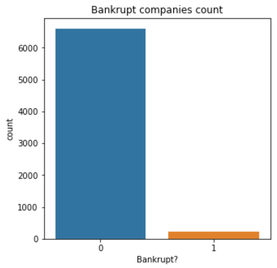

Figure 1 Count of labels

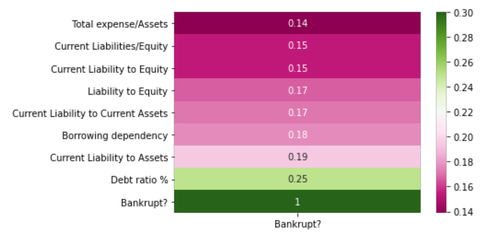

Figure 2 Co-relation with the target label

3. Selection of Models

3.1. K-Nearest Neighbors

- KNN algorithm is a supervised machine learning algorithm that can be used for both classification and regression scenarios. The working of the algorithm starts by calculating the closeness between the query and all the examples provided in the data. It then starts selecting the K number of examples that are close to the query. Then, it finishes by averaging the labels(regression) or ranking the most frequent label(classification).
- There are many ways of calculating the distance between the examples such as Cosine Distance, Jaccard Distance, Hamming distance and the most frequently used method Euclidean.
- KNN doesn't have a specialized training section and it uses all the data for training and hence called a lazy learning algorithm.

Advantages

- This is the simplest algorithm to implement [3]. It follows the memory-based approach which adapts to the new training data.
- It is quite flexible with the multi-class problem where it performs greatly without any extra effort.
- There isn't any training step in this algorithm, it tags the data points based on the historical data by choosing the majority class in the nearest neighbor.

Disadvantages

- This algorithm becomes slow when the size of the data gets increasing.
- It feels quite difficult to predict the output of the new data point as the number of variables increases.
- Choosing an optimal number of neighbors for consideration while classifying the new data point is the biggest issue.
- The performance of KNN on imbalanced data is very low.
- It doesn't have any solution to the missing value problem.

3.2. Naive Bayes

- Naive Bayes classifiers are nothing but a small collection of all the classification algorithms based on the Bayes Theorem.
- A common principle is shared between the family of classifiers which is every pair of features are independent of each other while classifying and this is called class conditional independence.

Advantages

- This algorithm performs well tested on the large data sets even with the training of small datasets
- It has a very low computational cost and can be used for multi-class prediction problems.
- This model is also termed as the best performing model on the text analytics kind of problems.
- It is a simple approach, an accurate and fast method in terms of prediction.
- This model performs well when compared to logistic regression or other models when the assumption of independence holds.

Disadvantages

- The major drawback of this algorithm is that it considers all the variables as independent which are contributing to the probability.
- One of the major disadvantages of this classifier is the Zero Probability problem [2]. It is nothing but for instance, if there is no training element of a particular class, the model is unable to make correct predictions.

4. Hypothesis Statement

- We are expecting that the Naïve Bayes algorithm performs much faster than the KNN due to real-time execution [5].
- As per the research paper[1], they have performed feature selection using various methodologies to determine the best variables for the models, however, I have chosen the top correlated features for my analysis. Also, to solve the problem of imbalance, I have used the Smoting technique, while the research paper was built on stratified sampling.

5. Methodology:

- The original data set is sampled by feature selection. It is divided into Training and Testing data sets by using HoldOut methodology with the help of cvpartition function.
- The test set is stored till the last for the testing of the data, while the training data is further slitted into 10 folds for the training using the KFold method.
- A model will be trained on the training dataset where the KFold assigns the folds for training randomly and iterates it along the loop to find the best fold. Once the model building is completed using the training folds, the testing fold will be predicted to check the performance of the model. The testing values keep changing along with the folds and best-trained split sets will be pushed to the final model. Thus, it acquires a better performance.
- Once the model is ready, it is tested on the original test dataset which is stored for the last step. After predicting the performance of the model on the test dataset, its accuracies are displayed, and confusion matrices will be filled to show its performance characteristics.
- The same methodology will be applied to the Naïve Bayes and testing performance will be derived.

6. Parameter selection and Experimental Results

6.1. KNN

- K Nearest Neighbors has performed well with the Smote data compared to the original data. It was able to predict the test labels very well as the data is balanced.
- A slight performance increase has been observed due to the normalization of the variables.
- Using the KFold method made it easy to train the model on the parameters with high accuracy.
- Compared to the dataset having all the columns, the model was able to build the necessary adjustments very quickly with the smote dataset.

Parameter selection

The value of K has been chosen as 7 and the number of folds as 10 as this has been a favorite choice as per the paper[1].

6.2. Naive Bayes

- Due to the process of Smoting the data, the model was able to perform with a vast increase in performance. This is because of the Zero Probability problem this model has.
- Even with the use of KFold, the model was able to understand the data divisions accurately increasing the accuracy of the test dataset.

Parameter selection

The 10-fold method has been used as per the paper [1].

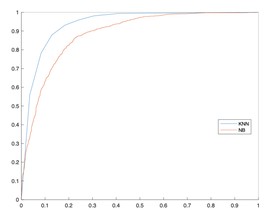

Figure 3 ROC Curve comparison

Bankruptcy Pridiction Results		
KNN	Model	**NB**
0.8805	Training Accuracy	0.6598
0.8762	Testing Accuracy	0.6362
0.8149	Recall	0.8069
0.9658	Precision	0.3352
0.8839	F1 Score	0.4736
0.9394	AUC	0.8922

Table 1 Prediction Results

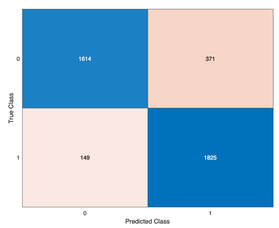

Figure 4 KNN Confusion Matrix

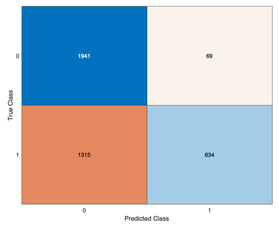

Figure 5 Naïve Bayes Confusion Matrix

7. Result from analysis and critical evaluation

- With the testing of this data set on KNN, the performance was accurate to a large extent. However, NB hasn't been performing up to the mark. This might be due to the low number of independent variables trained into the model.
- The Precision (Positive Predictive) value for the Naive Bayes has been the lowest as the model is detecting a huge number of False Positives which can be observed in Table 1. This might be due to the presence of some of the highly correlated features in the dataset which results in overinflating their importance [6].
- The ROC curve for both the models have shown a good result as shown in the Figure 3. KNN has been better than the NB for this dataset as the AUC value is greater for it. We can improve the AUC of the NB by considering a different feature selection method or no feature selection method. However, this costs the Precision and Recall of the model.
- As the number of features in the original dataset is huge, the run time for both the models have been significantly high. Due to the Smote technique and the correlation feature selection, the model was able to build quickly and test simultaneously with no time.
- Data normalization played a crucial role in the analysis as the Nearest Neighbours were able to generate easily and quickly with the increase of accuracy due to the normalized variables. This is primarily due to the ease of calculating the Euclidean distance between the data points which as widely spread on the axis scale.
- Considering 7 nearest neighbours has produced better results than other K-values which are tried based on Trial and Error. However, the K was chosen on the bias of the result from the research paper to determine more relevant results.
- When compared to the training of the model with the original dataset, the KFold method has increased the model building capacity. KNN model has increased its accuracy by an average of 10 per cent due to the implementation of the KFold method. This is due to the choice of a test set in the training phase for all the folds and reporting the mean of the observations as a result [7].
- As the result from the Naive Bayes are satisfactory, I have tried checking with other models at the end to clarify the performance in each method. I observed that few models are performing well on our final data set such as Random Forest and SVM. This can be further continued with the model parameters and obtain a few interesting results.

8. Lessons learned and future work

8.1. Lessons learned

- Using Hyperparameters in the KNN model has decreased the ability of the model to predict the outcome. This might be due to the large variables in the original dataset. However, with the correlated feature variables, KNN was able to improve a lot.
- Under sampling the data which has a huge imbalance is not a good choice as it decreases the data sample and the model built on this final dataset will not be robust. Compared to the dataset having all the columns, the model was able to build the necessary adjustments very quickly with the smote dataset.

8.2. Future Work

- T-test can be used for the critical feature selections as this is assumed to be the best method as per the results [1].
- Looking for all the factors that effect the business success and grouping the variables in this dataset and performing the analysis might produce better solutions
- ADASYN(Adaptive Synthetic Sampling) can be performed instead of SMOTE, and the result might have better improvements as this model performs data synthesis as per the data density

9. References

1. Deron Liang, C.-C. L. C.-F. T. G.-A. S., 2016. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research,* 252(2), pp. 561-572.
2. Navlani, A., 2018. *Naive Bayes Classification using Scikit-learn.* [Online] Available at: https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn [Accessed 4 12 2018].
3. Srivastava, T., 2018. *Introduction to k-Nearest Neighbors: A powerful Machine Learning Algorithm (with implementation in Python & R).* [Online] Available at: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/ [Accessed 26 3 2018].
4. Tsai, D. L. a. C.-F., 2020. *Taiwanese Bankruptcy Prediction Data Set.* [Online] Available at: https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction
5. Varghese, D., 2018. *Comparative Study on Classic Machine learning Algorithms.* [Online] Available at: https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222 [Accessed 6 12 2018].
6. Brownlee, J., 2014. Better Naive Bayes: 12 Tips To Get The Most From The Naive Bayes Algorithm. [Online] Available at: https://machinelearningmastery.com/better-naive-bayes/ [Accessed 10 12 2014].
7. Brownlee, J., 2020. Repeated k-Fold Cross-Validation for Model Evaluation in Python. [Online] Available at: https://machinelearningmastery.com/repeated-k-fold-cross-validation-with-python/ [Accessed 3 08 2020].