Methodologies for evaluating credit risk of business data using Machine Learning

Flinn Dolman*
flinn@brightnetwork.io

Version 1.0
https://www.brightnetwork.io

Abstract. In order to help address small businesses' lack of access to finance Bright is developing a credit risk model. To do so it is important that we develop a cogent and comprehensive approach to statistically analyse small business data. In this paper we explore what this means in the context of a system like Bright Network where there exists a means to utilise sophisticated operations on data. In doing so we outline our own experimental approach and how it was used to analyse borrower's credit risk indicators. Similarly we touch upon how our initial proof of concept risk model was constructed. We also discuss how advanced data utilisation techniques such as peer attestation as defined in our Data Vaults paper can be coupled with this proof of concept risk model. The ambition is that doing so will provide deeper insight into credit risk than one would typically expect given the quality of the data being analysed.

1 Introduction

Bright Network is a decentralised system built upon an open data sharing protocol. It allows small business data to be stored, verified and consumed for the purpose of accessing financial services. Its goal is to address the lack of access to finance these types of businesses face, in addition to scalability issues faced by lenders in the financial services sector. In our first technical paper describing Data Vaults[1], we introduced Autonomous Data: a standard for data and its utilisation. Its structural specification enables users of Bright Network to leverage their business data in new and unprecedented ways.

Data Vaults were designed to conform to the Autonomous Data standard. Accordingly, they exhibit its emergent utilities such as compatibility with linked data schemas and the ability to facilitate peer attestation. This makes them a fascinating specimen for statistical analysis. At Bright, we are developing a credit risk scoring service that performs such an analysis using Machine Learning techniques.

In this paper we will discuss how:

a. Machine Learning can be leveraged to assess credit risk.

b. New forms of data, including that which is unstructured and/or noisy, can be incorporated into credit risk modelling techniques to their benefit rather than detriment.

This paper will touch upon approaches in credit risk modelling and describe how they have informed our technical approach as well as our current credit risk model architecture.

*With thanks to Arnold Almeida and Jaime Van Oers

2 Credit Risk

Traditionally, assessment of credit risk was conducted by panels of domain experts based on several different characteristics. The most famous of these characteristics (known as the 5 C's[3]) were *character, capital, capacity, conditions* and *collateral*. The down side to this type of approach is that it is expensive (financially) and inefficient. Enter: Machine Learning.

2.1 Using Machine Learning

In recent times computational models have emerged as a powerful component to consider when assessing credit risk and deciding whether or not to lend to both consumers and institutions. Computational models allow investors to squeeze maximal returns out of their investments as well as providing a greater depth of information for use in risk management[5]. The models we will consider are created by providing data as input to Machine Learning algorithms. The end result is a system that given input in the form of loan applicant data can infer the probability of that applicant defaulting on a loan.

Typically, lending decisions made by Machine Learning models are based on Random Forests[6], Logistic Regression[7], Gradient Boosting[8] or Support Vector Machines[9]. These approaches have all shown they are able to perform well individually or as part of an ensemble when working in credit scoring domains[10].

2.1.1 Random Forests

Random Forests is a Machine Learning technique that makes predictions based on the outcome of a set of n Decision Trees votes. A Decision Tree is a tree-like flow chart that models decisions and their outcomes. The standout feature of Random Forests over other similar "bagging"[11] techniques is that it provides a means to generate trees that are de-correlated. This helps reduce the overall variance of the classifier. Random Forests generally performs well in both classification and regression tasks. An added benefit of this approach is that on a conceptual level Decision Trees are easy to understand and interpret, which allows for more insight into how the resulting model operates.

2.1.2 Gradient Boosting

Boosting strategies rely upon additively improving a model via inclusion of additional weak learners such as simple Decision Trees or Regressors. Gradient Boosting generalises this idea by allowing for optimisation of any loss function[12] so long as it is differentiable. Gradient Boosting is a popular approach in contemporary data science due to its relative efficiency compared to more complex approaches coupled with its typically impressive performance across a wide range of problem domains.

2.1.3 Logistic Regression

Regression is a class of statistical techniques whereby a relationship from input to output is determined. Logistic Regression maps an input to the range $[0, 1]$ which represents the classification probability of the input belonging to a specific class. The main pull of this technique is its relative simplicity both in computation, and interpretation of the resulting model.

2.1.4 Support Vector Machines

Support Vector Machines are models in which the optimal separating boundary (or hyperplane) is found between classes. The real power of Support Vector Machines is that they facilitate construction of non linear boundaries using a "kernel"[13]. By doing so, Support Vector Machines are able to tractably characterise many different trends in data.

2.2 Limitations

2.2.1 Bias in the data *(Class Imbalance)*

Large amounts of data are necessary to produce effective computational models for assessing credit risk. However, no matter the quantity of data available for analysis, one must be vigilant of whether the data is truly representative of the real world, rather than a cherry-picked subset. Specifically, we must account for inherent biases that exist within the data when performing analysis.

2.2.2 Limitations given the lack of quality data *(Signal to Noise)*

A central focus of Bright is to close the credit gap[4]. In order to do so our prerogative is to provide loans to companies that are typically overlooked by traditional lending institutions due to them having a lack of high quality verifiable data. As a result most of the features we are working with have a low signal to noise ratio.

2.3 Combating Limitations

2.3.1 Low Signal to Noise

There are many ways to address this issue. Most effective approaches will typically begin with exploratory data analysis (EDA). EDA allows an analyst to get a better understanding of the data, thus enabling the extraction of as much signal from the data that noise might be obfuscating. Feature importance analysis is also performed to construct new features according to observations of emergent trends in the data. These new features can be made to either amplify underlying signal or filter out noise. A major focus in designing our risk model was EDA and engineering new features.

An interesting point to note is that the ability to peer attest any data conforming to the Autonomous Data standard should theoretically help reduce noise itself. A considerable proportion of the noise we are addressing stems from the ambiguity of truths in user submitted data. The Attestation strategy proposed in the Data Vaults paper[1] is an effective tool to address this.

2.3.2 Class Imbalance

Class imbalance severely affects how you might grade a classifier's performance. For example, when classifying data into two categories a classifier might learn to always predict the class with the most entries in the training data set. Such a model might achieve a high classification success rate on test data, but in reality has poor predictive capability.

This problem is typical in the credit risk domain as it is likely there will be far more cases of a borrower not defaulting than defaulting. This is because lenders are unlikely to loan to borrowers who they believe are likely to default in the first place. Accordingly, our choice of performance metric is important.

3 Our Approach

We acknowledge that there is no general best way of analysing data and developing a risk model. In this sense, credit risk modelling is not a solved problem. Through our analysis we determined approaches that performed well for our dataset and iterated upon their design to find those which were best performing. This process was then repeated, each time eliminating the least effective designs given the performance metrics described in this section.

4 Performance Metrics

The domain in which we wish to train our model greatly informs the metrics that we use to grade its performance. For example, consider a model which determines whether a patient has cancer. In the ideal world we would like this model to have 0 false negatives and false positives. However, if we had to compromise we would rather misclassify someone without cancer than misclassify someone

with it. A similar argument can be made for lending: We would rather withhold lending to someone we misclassify as likely to default compared to lending to someone we misclassify as non-defaulting.

Given all considerations discussed we opted to evaluate performance according to two concepts that are largely supported in contemporary Machine Learning; AUC and Recall[14][15].

4.1 AUC

The area under the ROC curve (AUC)[16] is a relatively popular metric in Machine Learning. It factors in the true positive and false positive rate which allows it to honestly grade a classifier that is predicting on a dataset with a positively labeled minority class. In our analysis we used AUC as a general guidance metric to inform how successful our model iterations were. All that said, if one was to trust only the ROC and its derived AUC as a metric then they would risk having a somewhat optimistic perspective of the classifiers performance. This comes down to the fact that the False positive rate (the x axis of the ROC curve) does not drop drastically for large values of the denominator ($FP + TN$).

4.2 Recall

We used Recall[17] as a secondary metric to provide more insight and validate that our AUC scores were not giving us an unrealistic depiction of what was truly going on. As touched on before, we more highly value a classifier that occasionally misclassifies *non-defaults* as *defaults* than one that occasionally classifies *defaults* as *non-defaults*. The desired quantity which models this goal is *default* (class 1) Recall. The higher the *default* recall the more *defaults* we correctly identify. In order to avoid situations where best-performing models predict *defaults* 100% of the time, we balance our models by also monitoring *non-default* Recall.

In conclusion, our model selection criteria balances high *default* Recalls with *non-default* recalls that are significantly better than a uniform "dummy classifier"[18].

5 Training Data

Note: The purpose of the following sections is to detail our approach rather than to provide results.

Bright has access to a large database whose entries correspond loan applications made on behalf of small businesses.

The features in our dataset include:

a. Requested loan amount

b. Historical credit inquiries

c. Revenue

d. Loan purpose

6 Process

Our development of the risk model was augmented in rounds. Much of the flow within each round was similar, so for clarity key repeated processes are outlined here.

6.0.1 Feature engineering

As previously mentioned in Section 2.3.1, creative feature engineering on the data is required to reveal emergent trends and minimise its noise to signal ratio. Many such valuable features were created from seemingly arbitrary manipulations such as exponentiation.

6.0.2 Validation

Our typical process for validating model results had two components. Initially we would evaluate the effectiveness of a model by calculating relative frequency histograms of test set AUC over a large number of random splits of the dataset. Then we would determine cross validated *default* and *non-default* recalls for any models which exhibited low variance and high mean AUC scores. Constructing empirical distributions of AUC scores over random splits allowed us to perform statistical significance tests as well as apply other effective statistical methodologies that are seldom seen in typical Machine Learning workflows. This allowed us to make informed decisions about whether or not it was worth considering the resultant Recall scores of a model.

6.0.3 Hyper parameter tuning

Once an effective model architecture had been constructed, we optimised its hyper-parameters according to the resulting AUC by means of a grid search[19].

We considered alternative approaches to finding optimal hyper-parameters such as Bayesian optimisation [20], but found the marginal improvement in results didn't justify the added complexity they introduced.

6.1 Significance

Finally, we tested the significance of our best performing model architectures and their optimal hyper parameter choices. Our significance test involved comparing the AUC distribution generated by training on splits of the dataset against the AUC distribution of splits where the data had all features randomly shuffled.

7 Evaluating Machine Learning Models Independently

We started by looking to eliminate models that were performing poorly under a naive configuration. The AUC frequency results of models are shown in Figure 1.

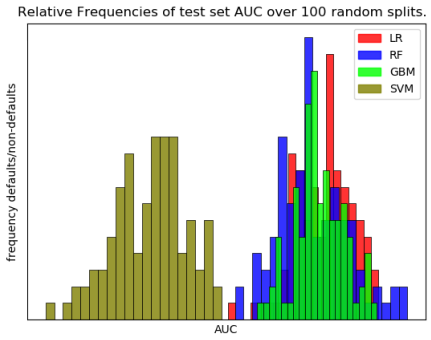

Figure 1: Histogram of AUC per test set split for different naively configured models.

7.0.1 Discarding Random Forests and Support Vector Machines

Inspecting the results in figure 1 allows the conclusion that SVMs were significantly outclassed by Logistic Regression and Gradient Boosting. Additionally, upon observation of resultant class 1 Recall for each model on an unseen validation set we noted that Random Forests were also outclassed by the other approaches. We decided to move on to the next stage of analysis without either.

7.0.2 Selecting Logistic Regression and Gradient Boosting

Redundant features were then removed by testing for co-linearity and mutual dependencies between features that were most important. An example of a model's relative performance before and after

this feature engineering is shown in Figure 2.

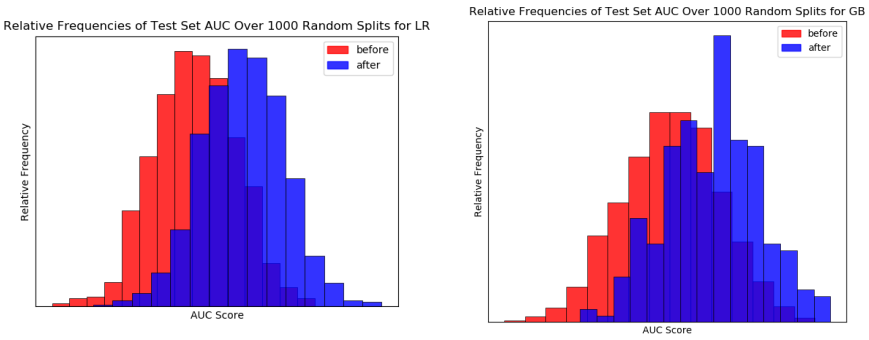

Figure 2: Histogram of AUC per test set split before and after feature engineering for Logistic Regression and Gradient Boosting.

8 Creating An Ensemble

During the analysis of different models we discovered that whilst they independently achieve a similar AUC, their predictions were rarely equivalent. This inspired us to start exploring ways in which we could combine models to produce an ensemble that performs better than any one of its constituents.

Figure 3 shows the performance of our individual best performing model, Logistic Regression, compared to the new ensemble model.

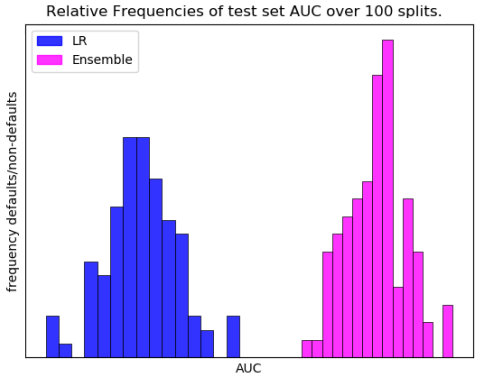

Figure 3: Histogram of AUC per test set split of a trained ensemble compared to the next best individual model; Logistic Regression.

8.0.1 Hyper parameter optimisation

Finally, optimising hyper-parameters of the individual models and the ensemble yields the most successful results. An example of these improvements are shown in Figure 4.

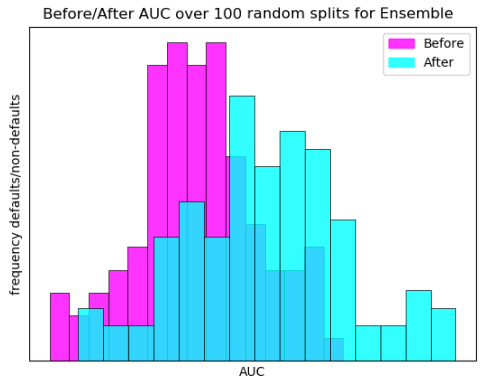

Figure 4: Histogram of AUC per test set split of the ensemble model before and after hyperparameter tuning.

8.0.2 Validation

As described in Section 6.1, we tested the significance of our model by training it with data where feature labels have been randomly shuffled. As expected, the end result is our ensemble model which on average produces an AUC that is considerably greater than the AUC of the same model trained on data with shuffled features.

The significance of our final model is shown in Figure 5.

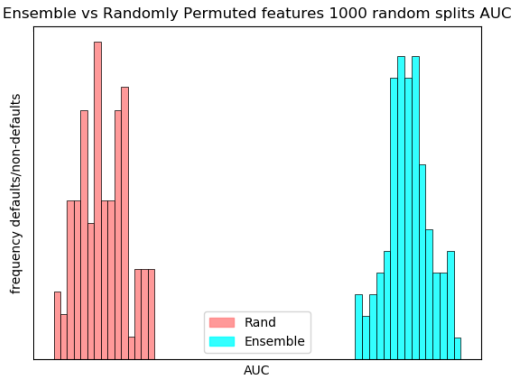

Figure 5: Histogram of AUC per test set split of the ensemble model trained with both normal test data and data where feature labels are randomly shuffled.

9 Conclusion

In this paper, we have outlined our approach for modelling small business credit risk given unstructured and noisy data. Its contents describe the process of constructing an ensemble of individual models and performing optimisations on both the architecture and our data in order to meet our contextual performance metrics. Finally we demonstrated the predictive significance of our approach when compared to the same model on completely randomised data.

9.1 Further Work

The modelling approach presented here represents the foundation of our credit risk scoring methodology based upon loan application data. Ongoing research continues to inform optimisations of our model architecture and enhance our feature engineering efforts. For example, we are exploring how we can reduce noise in our data by augmenting our data engineering with digital signal processing methodologies[21].

9.1.1 Emergent utilities of Data Vaults

Autonomous Data as described in our Data Vaults technical paper[1], creates possibilities for new dimensions of analysis in our risk model. One dimension we are particularly interested in is the statistical dependency of actions and interactions between agents in the network. Consideration of this factor will enable us to assess the likelihood an agent is engaging or will engage in certain actions on the network. This can then be used to determine the relative trustworthiness of an agent which will enable detection of malicious activities such as fraud or financial crime.

Additionally, the Data Vaults system facilitates peer attestation of business data. Whether or not a piece of data has been attested, and who that data has been attested by, could provide rich metadata and signals which allow more informed lending decisions to be made. The extensible nature of Data Vaults, and their open ended definition of how data can be utilised will continue to catalyse increasingly deep insight as the network grows. The dimensionality introduced by this insight will confer an exponential improvement in overall predictive capability of the models we go on to develop. These ideas will be elaborated upon in future technical papers.

References

[1] Data Vaults: Strategies for storing, verifying and sharing private data, `https://docsend.com/view/tfm4gns`

[2] Moody's Analytics: Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modeling.

https://www.moodysanalytics.com/risk-perspectives-magazine/managing-disruption/spotlight/machine-learning-challenges-lessons-and- opportunities-in-credit-risk-modeling

[3] GPA Capital: Five C's of Credit.
http://gp-assoc.com/wp-content/uploads/2017/10/5-Cs-of-Credit.pdf

[4] IFC Advisory Services: Closing the Credit Gap for Formal and Informal Micro, Small, and Medium Enterprises,
https://www.ifc.org/wps/wcm/connect/4d6e6400416896c09494b79e78015671/Closing+the+Credit+Gap+Report-FinalLatest.pdf

[5] Moody's Analytics: Maximize Efficiency: How Automation Can Improve Your Loan Origination Process
https://www.moodysanalytics.com/-/media/whitepaper/2017/maximize-efficiency-how-automation-can-improve-your-loan-origination-process.pdf

[6] Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

[7] Cox, D.R., 1958. The regression analysis of binary sequences. Journal of the Royal Statistical Society. Series B (Methodological), pp.215-242.

[8] Mason, L., Baxter, J., Bartlett, P.L. and Frean, M.R., 2000. Boosting algorithms as gradient descent. In Advances in neural information processing systems (pp. 512-518).

[9] Cortes, C. and Vapnik, V., 1995. Support-vector networks. Machine learning, 20(3), pp.273-297.

[10] Wang, G., Hao, J., Ma, J. and Jiang, H., 2011. A comparative assessment of ensemble learning for credit scoring. Expert systems with applications, 38(1), pp.223-230.

[11] Breiman, L., 1996. Bagging predictors. Machine learning, 24(2), pp.123-140.

[12] Wald, A., 1950. Statistical decision functions.

[13] Hofmann, T., Schölkopf, B. and Smola, A.J., 2008. Kernel methods in machine learning. The annals of statistics, pp.1171-1220.

[14] Hossin, M. and Sulaiman, M.N., 2015. A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), p.1.

[15] Sokolova, M. and Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), pp.427-437.

[16] Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters, 27(8), pp.861-874.

[17] Powers, D.M., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

[18] Sci-kit Learn Documentation, DummyClassifier
http://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html

[19] Chicco, D., 2017. Ten quick tips for machine learning in computational biology. BioData mining, 10(1), p.35.

[20] Mockus, J., Eddy, W. and Reklaitis, G., 2013. Bayesian Heuristic approach to discrete and global optimization: Algorithms, visualization, software, and applications (Vol. 17). Springer Science & Business Media.

[21] Vaseghi, S.V., 2008. Advanced digital signal processing and noise reduction. John Wiley & Sons.