*Workshops at 18th International Conference on Intelligent Environments (IE2022)*
*H.H. Alvarez Valera and M. Luštrek (Eds.)*

*doi:10.3233/AISE220039*

# Enhancing Loan Default Prediction with Text Mining

Barry EGAN [a,1] and Dr Kyle GOSLIN [b]
[a] *Technological University of Dublin*
[b] *Technological University of Dublin*

**Abstract.** Credit scoring is a popular method used by financial institutions to evaluate an applicants' risk of default. However, in certain circumstances, an individual's credit score is not an accurate indicator of their risk of default as it may be based on outdated information from a single point in time, or individuals may have no prior credit history from which to build the credit score. Several studies have investigated using text data to enhance the classification of loan default, with varying degrees of success. This research examines if the text data contained in the loan applications of a peer-to-peer (P2P) lending platform can be utilized to enhance loan default prediction. In this research, two models were created and optimized: one using only text data and the other using numeric data. The text and numeric models were then combined to see whether the classification performance of the individual models can be enhanced. The classification performance of the text model was superior to that of the numeric model, achieving accuracies 15.73% and 33.82% higher; however, by combining the models, there was a considerable improvement to the model's classification performance of between 2.8% and 19.87% respectively. Results showed that text data holds significant value for assessing credit risk, and when text data and numeric data are combined there is an enhancement in the prediction of loan default.

**Keywords.** Loan Default Prediction, Text Mining, Classification

## 1. Introduction

The recent introduction of innovative technologies has made radical changes to the financial services industry and its operations. The new generation banks, such as Revolut[2], and Peer to Peer Lenders (P2P), such as The Lending Club[3], have brought new technologies and modern approaches to banking, causing increased competition. This increased competition has resulted in traditional banks implementing and emphasizing modern technologies and services to keep pace and offer the level of customer service provided by these alternative banks and lenders (Jayasree and Vijayalakshmi Siva Balan, 2013).

The implementation of modern banking technologies, such as mobile banking, has resulted in more customer data being collected. Banks can use this data to their advantage to improve their decision-making process across several areas. However, traditional data

[2] https://www.revolut.com/en-IE
[3] https://www.lendingclub.com/

analytics is slow and is often unsuccessful as decisions are not accurate due to the complexity and amount of data being analyzed, resulting in hidden insights in the data not being unearthed (Zakirov and Momtselidze, 2015). Due to these issues, data mining has been embraced by the financial services sector to unearth significant insights from the customer data to allow for an enhanced decision-making process. Data mining has been applied in several areas in financial services, such as sentiment analysis, fraud detection, market segmentation, customer churn, and credit approval (Farooqi and Rashid Farooqi, 2017).

Credit scoring has been adopted as one of the core methods for financial institutions to gauge credit risk and aid in the decision-making process. Financial institutions can use credit scoring to decide to approve or reject an application for a loan (Keramati and Yousefi, 2011). Recently, there has been an increased focus on developing and utilizing machine learning techniques to analyze data to aid the credit decisions of financial institutions to better manage credit risk (Yap, Ong and Husain, 2011). Models being developed must consider both the financial loss associated with approving candidates who are at risk of default and the business loss of a candidate with a low risk of default being rejected for a loan (Markova, 2021).

As historical data belonging to an applicant is being used to make credit approval decisions, it can be difficult to make an accurate decision as the information is from a single point in time and may be outdated (Aphale and Shinde, 2020). Text data can help banks alleviate issues concerning this and, in turn, enhance their credit risk assessment, precisely their qualitative evaluation method. A large quantity of real-time textual data belonging to customers is available from social media, discussion boards, and chat rooms for financial institutions to utilize. If text mining techniques were applied to this data, informative credit risk insights could be obtained, which can complement insights gained from quantitative data analysis (Chen et al., 2017).

In this paper Section 2 describes previous research, Section 3 discusses the findings from the Data Exploration phase and the steps taken in preparing the data for the research. Section 4 discusses the steps taken to build the Text and Numeric models. Section 5 discusses the classification results of the models, the characteristics of non-defaulting and defaulting individuals, and the most predictive attributes, and Section 6 concludes this research outlining the core findings.

## 2. Literature Review

In research conducted by Guo et al. (2016), a two-tier ensemble model was utilized to predict loan applicants' creditworthiness based on the applicant's banking and social media data. In the research, the stacking model, the Tier-1 classifier, used low-level features as input, while the boosting-based ensemble model, the Tier-2 classifier, used low-level and high-level features. The Tier-1 stacking model consisted of a Decision Tree, Naïve Bayes, Logistic Regression, and Support Vector Machine model, while the Tier-2 classifier was a Gradient Boosted Decision Tree. The three low-level features used were based on demographic, tweet, and network features. The high-level features consisted of n-gram and topic distributions features and the predicted labels from the Tier 1 classifier. The research used two datasets, one using a balanced class distribution and the other an imbalanced class distribution. Combining the low-level features resulted in the best classification performance instead of using any single or combination of the features for both datasets. Similarly, combining low-level and high-level features resulted in the best classification performance for the models on both datasets. Combining low and high-level features resulted in the best-performing model overall in the research with an accuracy of 58.76% and 63.75% on the balanced and imbalanced datasets, respectively.

The research of Netzer, Lemaire and Herzenstein (2019) used the text data from loan applications and an ensemble model employing stacking to assess the creditworthiness of loan applicants from data acquired from a crowdfunding platform. Three tree-based models and two Logistic Regression models that employed differing forms of penalization were the stacking models were used in the research. The weightings for the stacking models were: Random Forest K-Best = 0.218, Random Forest Variance-Select = 0.116, Extra Trees = 0.066, Logistic Regression 1 = 0.040, and Logistic Regression 2 = 0.560." In the research, the stacking model was evaluated on text data only, financial/demographic data only, and text and financial/demographic data combined. Additionally, the ensemble was tested on three subsets of individuals: those with low credit grades, medium credit grades, and high credit grades. Across all grade subsets, financial data achieved superior results to text data; however, combining text and financial data resulted in the best performance. The combined data model achieved an AUC of 72.60% and a Jaccard Index score of 39.50%.

In a study by Niu, Ren and Li (2019), the researchers used P2P lending data to build a classification model for predicting loan default. In the study, researchers evaluated the classification performance of a Random Forest, AdaBoost, and LightGBM, a form of Gradient Boosted Decision Tree. The classification performance was based on using financial data only and when financial and text data were used for prediction. The performance of all three models improved across the three metrics used for evaluation of accuracy, AUC, and f1 score, when both financial and text data were used for classification. Overall, the best-performing model in the study was the LightGBM model. The LightGBM achieved an accuracy of 66.22%, an AUC of 71.1%, and an F1 score of 65.9%.

Based on previous studies in the area, several aspects helped to guide the research conducted. Stacking was implemented in the studies of Netzer, Lemaire, and Herzenstein (2019) and Guo et al. (2016) to combine the models effectively, and so was applied later in the research for combining the text and numeric models. Combining the power of numeric and text data resulted in an enhanced classification performance in the three studies reviewed. Finally, a Logistic Regression or Random Forest model was implemented in several studies and performed effectively in classifying loan default, so they would also be implemented in the research conducted.

For this research the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology was used.[4]

[4] https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview

## 3. Data Exploration and Preparation

### *3.1. Dataset*

The Lending Club, a Peer To Peer lending platform, provided the dataset for the study. The dataset consisted of 23 attributes (12 numerical, 11 categorical). The attributes were related to socioeconomic and demographic information regarding the applicant and loan-specific information. There were 119,945 records in the dataset, with 84% of records pertaining to fully paid loans and 16% defaulted loans.

### *3.2. Data Exploration*

In the dataset, 84.43% of the loan records related to fully paid loans, while 15.57% related to loans that had been charged off. 34.48% of records had a *B-loan grade*, 23.36% had a *C-grading*, and 20.22% of loans had an *A-grade*. *D*, *E*, *F*, and *G-grades* accounted for 13.10%, 5.71, 2.50%, and 0.63% respectively. The most common purpose a loan was used for was *debt consolidation*, which accounted for 56.98% of the records.

The *loan amount*, *annual income*, and *debt-to-income ratio* attributes were all correlated to the ability to pay. However, based on the exploration, the *credit score* and *interest rate* attributes held contained the most predictive power of the continuous attributes. As an individual's *credit score* rose, the default rate declined quickly. Similarly, as the interest rate rose, the default rate rose quickly. The monthly *installments* an individual was required to repay did not appear to bear any evident influence on the default rate amongst individuals.

The *purpose* for which an individual was using the loan impacted the default rate amongst individuals. A loan used for financing *business activities* or *renewable energy* had 27% and 20% default rates, respectively, making these loans the riskiest. In comparison, loans used for *cars* or *major purchases* held the lowest risk, with default rates of 9.8% and 10.71%, respectively. The *term* of the loan was closely linked to the default rate, with shorter loan terms considerably less risky. A *60-month* term loan had a default rate of 28.33%, while a *36-month* term loan had a 12.22% default rate, a reduction of 16.11%.

However, a loan's *grade* and *sub-grade* held the most predictive power of the categorical attributes. A loan's *grade* and *sub-grade* are based on an in-house calculation based on the applicant's *credit score*, *annual income*, *debt-to-income ratio*, and loan-specific information such as the *purpose*, *term*, and *loan amount* to calculate the *grading*. The *gradings*, in turn, decide interest rates, so they possess strong predictive power to whether an individual will default or not. As the *grade* given to an individual's loan falls, the default rate rises rapidly.

### *3.3. Data Preparation*

The first step in data preparation was to transform the *loan grade*, *loan sub-grade*, *term*, and *state* attributes. The *term* attribute was a string; however, it was converted to a number for the research. As the research may use a text pre-processing technique that filters out values based on length, it was decided to add the *term* "*GradedLoan*" to the end of the loan *grade* and *sub-grade* to ensure they were kept in the wordlist due to their predictive power, and not removed in error. Additionally, to ensure that the linguistic structure of the *state* value was maintained, any spacing between words in the *state* attribute was removed, so "*New York*" was now "*NewYork*".

Variance inflation factor was then used to measure the level of multicollinearity amongst the continuous variables in the dataset. However, there was no evidence of multicollinearity detected. In several previous pieces of research in the area, such as Wang et al. (2016), an individual's education level was said to possess predictive power in regard to predicting default. For this reason, it was decided to create a *readability score* attribute based on the words used in the *description* attribute. A *readability score* is calculated based upon the diversity of vocabulary and complexity of sentences in a piece of text. A *readability score* reflects an individual's cognitive ability and is a good determinant of education level.

The next step was to create the training and test datasets. Initially, a training dataset consisting of 4,000 records and a test set with 1,000 records, was created. The training dataset had a balanced class distribution, and the test dataset reflected the class imbalance of the dataset. However, later in the research, it was desired to test the model's performance on larger training and test datasets. The larger training and test datasets were created using records not used in the original training and test datasets. The training dataset consisted of 30,000 records, and the class distribution was balanced, the test dataset consisted of 10,000 records, and the class distribution was imbalanced, the same as the original dataset.

Outlier detection was performed using Mahalanobis Distance from the Scipy[5] library in Python on the training dataset. Using a confidence interval of 0.99, the outlier detection detected 74 potential outliers. However, on closer inspection, it was deemed that these outliers are not truly outliers and naturally differ from the average record.

## 4. Model Building

The research used an iterative approach whereby as the model build progressed through each phase, the pre-processing techniques for each model that demonstrated the optimum performance were retained as the benchmark for each subsequent step in the process to be measured against. Once the research had run through the pre-processing steps, the model was optimized by re-evaluating different operators to ensure the final model was as strong as possible and as many different implementations of the models were evaluated.

The text models evaluated were Naïve Bayes, Support Vector Machines, and a Decision Tree. The Naïve Bayes model was the best performing of the three models evaluated, so it was used for the text model. The numeric models evaluated were Random Forest and Logistic Regression. The Logistic Regression model was the best performing of the two

[5] https://scipy.org/

models evaluated, so it was used for the numeric model. The models in the study were built using Rapidminer Studio 9.10[6].

Stacking was used for combining the models. The optimum Naïve Bayes and Support Vector Machine text models were the base learners, and the optimum Logistic Regression numeric model was the stacking model learner.

### *4.1. Text Model*

For the text model, the proposed Naïve Bayes model consisted of the following steps:

**Text Parsing:** Non-letters tokenization was used whereby each word in a text is a token, but symbols and numbers are excluded from acting as tokens. Additionally, transforming text to lower cases, identifying synonyms, and four stemming operators were investigated, namely, the Porters, Lovins, Snowball, and WordNet stemmers; however, none of these improved the performance of the text model.

**Text Filter**: Stopwords were filtered out for the text model to remove any common words occurring in a high frequency across the texts. When a word repeatedly occurs in documents, it results in the creation of noise in the data. These repeatedly occurring words have minimal benefit to the overall model and should be removed or filtered out. The filtering of tokens based on a minimum and maximum character length was also employed. Tokens with less than two characters and over 12 characters were filtered out for the text model. The implementation of pruning of words based on the occurrence of a token over and under different thresholds was investigated. However, no implementation improved the text models classification performance.

**Feature Extraction:** The Weight by SVM operator was implemented to calculate the importance of the terms in the wordlist for the final model. The Select by Weight operator was implemented to determine the optimal number of terms for the final model. The number of terms was set to a range of values from 1,000 to 20,000 terms, with 16,000 terms resulting in the best classification performance. Tri-grams were generated for the final model, resulting in the model's best classification performance. The implementation of Unigrams, Bigrams, and Four-grams was investigated but implementing Tri-grams resulted in the optimum performance. Four techniques were evaluated to create word vectors: Term Frequency – Inverse Document Frequency (TF-IDF), Term Frequency, Term Occurrences, and Binary Term Occurrences. The implementation of TF-IDF resulted in the best classification performance.

### *4.2. Numeric Model*

The proposed Logistics Regression numerical model used standardization, meaning that numeric columns had zero mean and unit variance. An intercept was added, resulting in a constant term being included in the model. Additionally, p-values were computed for the model, and colinear columns were removed. Mean Imputation was used to deal with missing values for the model. The maximum number of iterations for the model was set to 1, while no maximum runtime was implemented for training the model.

## 5. Evaluation of Results

### *5.1. Classification Results*

**Table 1.** Optimum Text & Numeric Models Classification Performance - Training Dataset - 4,000 Records

| Training Dataset – 4,000 Records | Fully Paid | | | Charged Off | | | Overall Accuracy | STD DEVIATION |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 Score | Precision | Recall | F1 Score | | |
| Optimum Text Model - Naïve Bayes | 98.39% | 98.05% | 98.22% | 98.06% | 98.40% | 98.23% | 98.22% | 0.79% |
| Optimum Numeric Model - Logistic Regression | 63.79% | 66.60% | 65.16% | 65.06% | 62.20% | 63.60% | 64.40% | 2.15% |

**Table 1** shows the classification performance of the Naive Bayes text model and Logistic Regression numeric model on the training dataset consisting of 4,000 records. The performance of the Naïve Bayes text model is significantly superior to that of the Logistic Regression numeric model across the metrics evaluated. The Naïve Bayes text model has an F1 score on the Fully Paid class of 98.22% and the Charged Off class of 98.23%, 33.05%, and 34.63% better than the Logistic Regression numeric model, which had F1 scores of 65.16% on the Fully Paid class and 63.60% on the Charged off Class.

The Naive Bayes text model had an accuracy of 98.22%, which was 33.82% better than the Logistic Regression numeric model's accuracy of 64.40%. Additionally, the Naive Bayes text model was more stable than the Logistic Regression numeric model with a standard deviation of 0.79%, 1.36% lower than the Logistic Regression numeric model's standard deviation of 2.15%.

**Table 2.** Optimum Text , Numeric & Combined Models Classification Performance - Testing Dataset - 1,000 Records

| Test Dataset – 1,000 Records | Fully Paid | | | Charged Off | | | Overall Accuracy |
|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 Score | Precision | Recall | F1 Score | |
| Optimum Text Model - Naïve Bayes | 100.00% | 92.14% | 95.91% | 70.80% | 100.00% | 82.90% | 93.40% |
| Optimum Numeric Model - Logistic Regression | 84.33% | 99.29% | 91.20% | 45.45% | 3.12% | 5.84% | 83.90% |
| Combined Models | 98.78% | 96.67% | 97.71% | 84.27% | 93.75% | 88.76% | 96.20% |

**Table 2** shows the classification performance of the Naive Bayes text, Logistic Regression numeric, and Combined models on the testing dataset consisting of 1,000 records. The performance of the Combined model is slightly superior to that of the Logistic Regression numeric model overall but significantly superior on the Charged Off class across the metrics evaluated. The Combined model is slightly superior to the Naive Bayes text model across the metrics evaluated. The Combined model has an F1 score on the Fully Paid class of 97.71%, 1.80% better than the Naive Bayes text model's F1 score of 95.91%, and 6.51% better than the Logistic Regression numeric model's F1 score of

91.20% for the Fully Paid class. The Combined model has an F1 score on the Charged Off class of 88.76%, 5.85% better than the Naive Bayes text models F1 score of 82.90%, and 82.92% better than the Logistic Regression numeric model's F1 score of 5.84%, for the Charged Off class.

The Combined model had an accuracy of 96.20%, which was 2.80% and 12.30% better than the Naive Bayes text and Logistic Regression numeric models' accuracies of 93.40% and 83.90%, respectively.

**Table 3.** Optimum Text & Numeric Models Classification Performance - Training Dataset - 30,000 Records

| **Training Dataset – 30,000 Records** | Fully Paid | | | Charged Off | | | Overall Accuracy | STD DEVIATION |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 Score | Precision | Recall | F1 Score | | |
| Optimum Text Model - Naïve Bayes | 88.17% | 71.09% | 78.71% | 75.78% | 90.46% | 82.47% | 80.77% | 0.63% |
| Optimum Numeric Model - Logistic Regression | 64.13% | 68.25% | 66.13% | 66.07% | 61.83% | 63.88% | 65.04% | 0.73% |

**Table 3** shows the classification performance of the Naive Bayes text model and Logistic Regression numeric model on the training dataset consisting of 30,000 records. The performance of the Naïve Bayes text model is considerably superior to that of the Logistic Regression numeric model across the metrics evaluated. The Naïve Bayes text model has an F1 score on the Fully Paid class of 78.71% and the Charged Off class of 82.47%, 12.59%, and 18.59% better than the Logistic Regression numeric model, which had F1 scores of 66.13% on the Fully Paid class and 63.88% on the Charged off Class.

The Naive Bayes text model had an accuracy of 80.77%, which was 15.73% better than the Logistic Regression numeric model's accuracy of 65.04%. Additionally, the Naive Bayes text model was more stable than the Logistic Regression numeric model with a standard deviation of 0.73%, 0.10% lower than the Logistic Regression numeric model's standard deviation of 0.73%.

**Table 4.** Optimum Text , Numeric & Combined Models Classification Performance - Testing Dataset - 10,000 Records

| **Test Dataset - 10,000 Records** | Fully Paid | | | Charged Off | | | Overall Accuracy |
|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 Score | Precision | Recall | F1 Score | |
| Optimum Text Model - Naïve Bayes | 100.00% | 63.95% | 78.01% | 34.57% | 100.00% | 51.38% | 69.72% |
| Optimum Numeric Model - Logistic Regression | 84.61% | 98.70% | 91.11% | 45.77% | 5.75% | 10.22% | 83.33% |
| Combined Models | 91.85% | 96.14% | 93.95% | 73.16% | 55.19% | 62.92% | 89.59% |

**Table 4** shows the classification performance of the Naive Bayes text, Logistic Regression numeric, and Combined models on the testing dataset consisting of 10,000 records. The performance of the Combined model is slightly superior to the Naive Bayes text and Logistic Regression numeric models across the metrics evaluated. The Combined model has an F1 score on the Fully Paid class of 93.95%, 2.83% better than the Logistic Regression numeric model's F1 score of 91.11%, and 15.93% better than the Naïve Bayes text model's F1 score of 78.01% for the Fully Paid class. The Combined model has an F1 score on the Charged Off class of 62.92%, 11.54% better than the Naive Bayes text model's F1 score of 51.38%, and 52.70% better than the Logistic Regression numeric model's F1 score of 10.22%, for the Charged Off class.

The Combined model had an accuracy of 89.59%, which was 6.26% and 19.87% better than the Logistic Regression numeric and Naive Bayes text models' accuracies of 83.33% and 69.72%, respectively.

The models in the research have achieved a superior classification performance to comparable models from studies in the area that used a similar-sized dataset. In research conducted by Niu, Ren, and Li (2019), the optimum performing Light GBM model achieved an accuracy of 66.2% and an average F1 score of 65.90%, using both text and numeric data on an imbalanced dataset, similar to this research. However, the closest comparison is to the stacking model used in research conducted by Guo et al. (2016). In this research, the optimum performing model utilizing text and numeric data achieved an accuracy of 58.76%. The classification performance of the models in these two studies is considerably inferior to the classification performance of the combined model in this study. The combined model in the research achieved an Accuracy and average F1 Score of 96.20% and 93.24% on the original test dataset and 89.59% and 78.44% on the extended test dataset.

### *5.2. Discussion*

In the research, the optimized numeric model used six attributes. Based on the weightings assigned by the Weight by SVM operator, the loan's interest rate was the most significant indicator of default, and the term was the second most important indicator of loan default. The year a loan was issued, the issue date of a loan, the income of an individual, and the debt-to-income ratio of the individual were also among the most important loan default indicators. Based on the earlier explorations findings of the research, the interest rate, term, and annual income were all correlated with the ability to repay; however, the issue date, the year the loan was issued, and the debt-to-income ratio showed no predictive power.

## 6. Conclusion

This paper has investigated the merit of applying text mining techniques to the textual information supplied in loan applications to enhance the loan default prediction. The merit of combining the Naive Bayes text and Logistic Regression numeric models is evident from the performance of the combined models on the two test datasets. By combining the models, the classification of both fully paid and charged-off loans improved considerably.

Regarding the models' classification performance, the models constructed in the research compared favorably to models from previous studies. Similar to the prior studies reviewed, combining the text and numeric models into a single model improved the classification performance. Additionally, implementing n-grams in the research resulted in a noticeable improvement to the text models' classification performance, as seen in Guo et al. (2016).

## 7. Future Work

Future research into the area will seek to optimize the model for use on larger datasets. The initial training dataset that the optimized models were constructed using was only 4,000 records due to computational restrictions. The exceptional results of the optimized text model and combined models on the smaller test dataset and the promising results achieved on the larger test dataset would indicate the potential for significant improvements if the model was constructed and optimized using a larger training dataset. Additionally, the research will investigate the merit of using different techniques for combining the text and numeric models to see if the classification performance improves.

## References

[1] Aphale, A. and Shinde, S. (2020). Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval. International Journal of Engineering Research & Technology (IJERT), [online] 9(8), pp.991–995.

[2] Chen, Y., Rabbani, R.M., Gupta, A. and Zaki, M.J. (2017). Comparative text analytics via topic modeling in banking. 2017 IEEE Symposium Series on Computational Intelligence (SSCI). [online] Available at: https://ieeexplore.ieee.org/document/8280945 [Accessed 30 Mar. 2021].

[3] Farooqi, R. and Rashid Farooqi, M. (2017). Effectiveness of Data mining in Banking Industry: An empirical study. International Journal of Advanced Research in Computer Science, 8(5), pp.827–830.

[4] Guo, G., Zhu, F., Chen, E., Liu, Q., Wu, L. and Guan, C. (2016). From Footprint to Evidence. ACM Transactions on the Web, 10(4), pp.1–38.

[5] Jayasree, V. and Vijayalakshmi Siva Balan, R. (2013). A REVIEW ON DATA MINING IN BANKING SECTOR. American Journal of Applied Sciences, 10(10), pp.1160–1165.

[6] Keramati, A. and Yousefi, N. (2011). A Proposed Classification of Data Mining Techniques in Credit Scoring. [online] Available at: http://www.iieom.org/ieom2011/pdfs/IEOM061.pdf [Accessed 29 Mar. 2021].

[7] Markova, M. (2021). Credit card approval model: An application of deep neural networks. SEVENTH INTERNATIONAL CONFERENCE ON NEW TRENDS IN THE APPLICATIONS OF DIFFERENTIAL EQUATIONS IN SCIENCES (NTADES 2020), [online] pp.1–7.

[8] Netzer, O., Lemaire, A. and Herzenstein, M. (2019). When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications. Journal of Marketing Research, 56(6), pp.960–980.

[9] Niu, B., Ren, J. and Li, X. (2019). Credit Scoring Using Machine Learning by Combing Social Network Information: Evidence from Peer-to-Peer Lending. Information, 10(12), p.397.

[10] Raju, S., Bai, R. and Chaitanya, K. (2014). Data mining: Techniques for Enhancing Customer Relationship Management in Banking and Retail Industries. International Journal of Innovative Research in Computer and Communication Engineering (An ISO, [online] 3297(1).

[11] Wang, S., Qi, Y., Fu, B. and Liu, H. (2016). Credit Risk Evaluation Based on Text Analysis. International Journal of Cognitive Informatics and Natural Intelligence, 10(1), pp.1–11.

[12] Yap, B.W., Ong, S.H. and Husain, N.H.M. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. Expert Systems with Applications, [online] 38(10), pp.13274–13283.

[13] Zakirov, D. and Momtselidze, N. (2015). Application of Data Mining in the Banking Sector. Journal of Technical Science and Technologies, 4(1).