

Statistical Methods in Credit Risk Prediction: Analyzing Risk through Data Analytics

Yuhui Wu

University of California-Irvine, 260 Aldrich Hall Irvine, CA 92697-1075, US

Abstract. This study explores the factors influencing loan approval decisions by financial institutions using logistic regression analysis. By examining a dataset of loan applicants, we identify key predictors that significantly impact the likelihood of loan approval. The analysis reveals that higher income, larger family size, higher education levels, possession of a CD account, use of online banking, and having a credit card are positively associated with loan approval. Conversely, the presence of a securities account has a more complex relationship with loan approval likelihood. These findings suggest that financial institutions favor applicants with stable financial profiles and higher levels of financial engagement. Understanding these factors can help improve risk assessment models, ensuring more accurate and fair loan approval processes.

Keywords: Credit Risk Prediction, Logistic Regression, Loan Approval, Big Data Analytics, Financial Risk Management.

1. Introduction

The financial sector faces a persistent challenge in extending credit to individuals with insufficient or non-existent credit histories. This issue becomes even more critical when potential borrowers exploit their lack of credit history, increasing the risk of default. According to recent statistics, approximately 45% of loan applicants in emerging markets have thin or no credit files, complicating the risk assessment process for financial institutions (World Bank, 2020). This high-risk segment often includes young adults, recent immigrants, and individuals who prefer to operate on a cash basis, presenting a unique challenge for lenders.

Thesis Statement: This paper explores integrating advanced statistical methods and big data analytics to enhance credit risk prediction, aiming to minimize financial institutions' exposure to defaults while optimizing loan approval processes.

Background: Background: Credit risk assessment has historically relied on traditional methods such as credit scoring models, primarily based on historical data on loan repayments. However, these models often fail to capture the nuanced financial behavior of individuals with limited credit history. The advent of big data analytics provides an opportunity to enhance these traditional models by incorporating alternative data sources such as social media activity, utility payments, and other non-traditional financial behaviors.

Purpose of Study: This study aims to evaluate the effectiveness of logistic regression models in predicting credit risk, using a comprehensive dataset from a banking institution. By analyzing various predictors, the research seeks to identify significant factors that influence loan approval decisions and suggest improvements in current risk assessment practices.

2. Literature Review

2.1. Big Data in Credit Risk Management

Big data analytics (BDA) has become integral to modern risk management frameworks, particularly in the financial sector. The literature emphasizes three primary characteristics of big data—volume, velocity, and variety—that necessitate innovative analytical techniques (Sagiroglu & Sinanc, 2013). Handling large datasets, rapid data processing, and diverse data sources is crucial for deriving meaningful insights for risk prediction (Hasnat, 2018).

Historical Development: The concept of big data emerged in the late 20th century, with significant advancements in data storage and processing capabilities. The financial sector quickly adopted these technologies to improve risk assessment and management. Early adopters of BDA in banking saw improvements in fraud detection, customer segmentation, and personalized financial services.

Modern Applications: Today, financial institutions use BDA to analyze transaction data, social media interactions, and other unstructured data sources. This holistic approach allows for a more accurate assessment of an individual's creditworthiness. Studies have shown that incorporating non-traditional data can significantly reduce default rates and increase the accuracy of credit scoring models (Chen et al., 2015).

2.2. Statistical Methods in Risk Prediction

Statistical methods such as logistic regression are pivotal in credit risk prediction. These models estimate the probability of a binary outcome (e.g., loan default) based on predictor variables. The methodology involves fitting a model to the data, evaluating its statistical significance, and interpreting the coefficients to understand the impact of various predictors on the likelihood of default (Vasarhelyi, Kogan, & Tuttle, 2015).

Logistic Regression: Logistic regression is a widely used statistical method in credit risk prediction due to its simplicity and interpretability. It models the relationship between a binary dependent variable and one or more independent variables by estimating probabilities using a logistic function.

Other Statistical Methods: While logistic regression is prevalent, other methods such as decision trees, random forests, and support vector machines are also used in credit risk prediction. These methods often provide higher accuracy but at the cost of interpretability. Recent advancements in machine learning have led to the development of ensemble methods that combine multiple models to improve prediction performance.

Comparison of Methods: Studies comparing different statistical methods have found that while complex models often outperform simpler ones in terms of accuracy, they require more computational resources and expertise. Logistic regression remains a popular choice due to its balance between simplicity and performance (James et al., 2013).

3. Methodology

3.1. Data Collection and Preparation

The dataset used in this study was obtained from a banking institution, focusing on personal loan applications. It includes several key variables crucial for credit risk assessment: (By the Variable description provided on Kaggle)

Age: Customer age.

Experience: Customer experience in years.

Income: Income of the customer.

ZipCode: Customer's residence zipcode.

Family: Number of family members of the customer.

CCAvg: Credit card average spending.

Education: Education level of the customer.

Mortgage: Whether the customer has a mortgage or not.

Personal Loan: Indicator if a personal loan was given (0 = No, 1 = Yes).

Securities Account: Whether the customer has a securities account.

CD Account: Whether the customer has a certificate of deposit (CD) account.

Online: Whether the customer uses online banking.

Credit Card: Whether the customer has a credit card.

3.2. Logistic Regression Analysis

The logistic regression model was fitted using JMP software (filtered with P-value that smaller than predetermined alpha, 0.05), focusing on the following predictors:

Income
Family
Education
Securities Account
CD Account
Online
Credit Card

The analysis aimed to identify significant predictors of loan approval status (Personal Loan: 0 = No, 1 = Yes).

Steps in Analysis:

1. **Data Cleaning:** Removed rows with missing values to ensure the integrity of the analysis.
2. **Variable Selection:** Selected relevant predictors based on the JMP fit Model function.
3. **Model Fitting:** Fitted the logistic regression model using JMP software.
4. **Model Evaluation:** Assessed the model's fit using metrics like Chi-square, R-squared, and AICc.

Data Preprocessing: Before fitting the model, the data was preprocessed to handle missing values and outliers. Missing values were corrected using the median for numerical variables and the mode for categorical variables. Outliers were identified using the Z-score method and were removed to reduce their impact on the model.

4. Results

4.1. Model Fit and Significance

The whole model test indicated a significant fit with a Chi-square value of 1864.994 ($p < 0.0001$), demonstrating the model's effectiveness in predicting loan approval status. Key metrics such as the R-squared value (0.5898) and AIC (1313.08) support the model's robustness and relative quality.

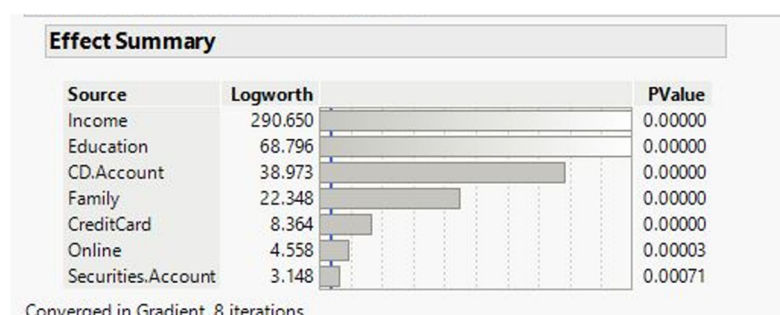
Model Coefficients: The logistic regression coefficients provide insights into the relationship between predictors and the likelihood of loan approval. For example, a positive coefficient for Income indicates that higher incomes are associated with a higher likelihood of approval.

Model Diagnostics: Various diagnostic measures were used to assess the model's fit, and the results indicated no significant lack of fit. Additionally, the model's residuals were analyzed to check for patterns that might suggest model misspecification.

4.2. Significant Predictors

Effect Summary:

The analysis identified several significant predictors of loan approval status. The "Effect Summary" chart below illustrates the Logworth and P-values for each predictor, indicating their statistical significance.



Description Of Effect Summary Chart: The "Effect Summary" chart shows the Logworth and P-values for each predictor variable in the logistic regression model. Logworth is the negative logarithm (base 10) of the P-value, providing a measure of the statistical significance of each predictor. Higher Logworth values indicate greater significance. (With smaller p-value, we have stronger evidence against the null hypothesis.)

Income: Significantly impacts loan approval status with the highest Logworth value of 290.650 and a p-value of 0.0000.

Education: Also significantly influences approval likelihood, with a Logworth value of 68.796 and a p-value of 0.0000.

CD Account: The presence of a CD account is a strong predictor with a Logworth of 38.973 and a p-value of 0.0000.

Family: The number of family members significantly affects loan approval, with a Logworth of 22.348 and a p-value of 0.0000.

Credit Card: Ownership of a credit card is another significant predictor, with a Logworth of 8.364 and a p-value of 0.0000.

Online: Using online banking services also significantly impacts approval status, with a Logworth of 4.558 and a p-value of 0.00003.

Securities Account: Having a securities account is statistically significant, with a Logworth of 3.148 and a p-value of 0.00071.

These results highlight the critical factors that influence loan approval decisions. Higher income, higher education levels, and the presence of financial products like CD accounts and credit cards are positively associated with loan approval. Additionally, the number of family members and the use of online banking services are significant predictors of loan approval likelihood.

Parameter Estimates:

The "Parameter Estimates" table provides detailed information on the estimated coefficients for each predictor variable, their standard errors, Chi-square values, and p-values. This information is crucial for understanding the strength and direction of the relationship between each predictor and the loan approval status.

Parameter Estimates				
Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-12.556781	0.5452492	530.35	<.0001*
Income	0.05739829	0.0024458	550.73	<.0001*
Family	0.69093737	0.0739287	87.35	<.0001*
Education	1.67556497	0.110966	228.00	<.0001*
Securities.Account[0]	0.45736264	0.1418079	10.40	0.0013*
CD.Account[0]	-1.9342931	0.1611774	144.02	<.0001*
Online[0]	0.3248619	0.0779788	17.36	<.0001*
CreditCard[0]	0.5633672	0.1022332	30.37	<.0001*

For log odds of 1/0

Description of Parameter Estimates:

Intercept: The intercept term has a significant negative coefficient (-12.556781), indicating the baseline log-odds of loan approval when all predictors are at zero.

Income: The coefficient for income (0.05739829) is positive and highly significant (Chi-square = 550.73, $p < 0.0001$), indicating that higher income increases the likelihood of loan approval.

Family: The coefficient for the number of family members (0.69093737) is positive and significant (Chi-square = 87.35, $p < 0.0001$), suggesting that having more family members is associated with a higher likelihood of loan approval.

Education: The coefficient for education (1.67556497) is highly positive and significant (Chi-square = 228.00, $p < 0.0001$), indicating that higher education levels greatly increase the likelihood of loan approval.

Securities Account: The coefficient for having a securities account (0.45736264) is positive and significant (Chi-square = 10.40, $p = 0.0013$), suggesting that having a securities account increases the likelihood of loan approval.

CD Account: The coefficient for having a CD account (-1.9342931) is negative and significant (Chi-square = 144.02, $p < 0.0001$), indicating a strong negative relationship with loan approval.

Online: The coefficient for using online banking (0.3248619) is positive and significant (Chi-square = 17.36, $p < 0.0001$), suggesting a positive relationship with loan approval.

Credit Card: The coefficient for having a credit card (0.5633672) is positive and significant (Chi-square = 30.37, $p < 0.0001$), indicating that having a credit card increases the likelihood of loan approval.

Ranking and Comparison of Predictor Influence:

Statistically Significant Influence (Parameter Estimates)

1. CD Account (Estimate = -1.9342931): The negative estimate for having a CD account indicates a strong negative influence on loan approval, with a substantial Chi-square value (144.02). This suggests that applicants with a CD account are less likely to be approved for a loan.

2. Education (Estimate = 1.67556497): The highly positive estimate and substantial Chi-square value (228.00) indicate that higher education levels are a crucial factor in loan approval. Applicants with higher education levels are significantly more likely to be approved for a loan.

3. Family (Estimate = 0.69093737): The estimate for family size suggests a strong positive impact on loan approval, supported by a significant Chi-square value (87.35). Applicants with larger families are more likely to be approved for a loan.

4. Credit Card (Estimate = 0.5633672): The positive estimate indicates a significant positive influence on loan approval, with a high Chi-square value (30.37). Applicants with a credit card are more likely to be approved for a loan.

5. Securities Account (Estimate = 0.45736264): The positive estimate shows that having a securities account positively influences loan approval, as indicated by a lower but significant Chi-square value (10.40). Applicants with a securities account are more likely to be approved for a loan.

6. Online Banking (Estimate = 0.3248619): The positive estimate suggests that using online banking services increases the likelihood of loan approval, supported by a significant Chi-square value (17.36). Applicants who use online banking are more likely to be approved for a loan.

7. Income (Estimate = 0.05739829): The positive estimate confirms that higher income increases the likelihood of loan approval. The high Chi-square value (550.73) reinforces its practical importance. Applicants with higher incomes are more likely to be approved for a loan.

Practically Significant Influence (Odds Ratios)

The analysis of odds ratios provides a clear understanding of the practically significant influence of each predictor on the likelihood of loan approval. The odds ratios quantify the change in odds of loan approval associated with a one-unit increase in the predictor variable, holding all other variables constant.

Odds Ratios				
For Personal.Loan odds of 1 versus 0				
Unit Odds Ratios				
Per unit change in regressor				
Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Income	1.059078	1.054013	1.064167	0.9442179
Family	1.995585	1.726403	2.306739	0.5011061
Education	5.341812	4.297681	6.639617	0.1872024
Securities.Account	0.400627	0.229787	0.69848	2.4960894
CD.Account	47.87465	25.45169	90.05225	0.0208879
Online	0.52219	0.38466	0.708892	1.9150118
CreditCard	0.32409	0.217082	0.483846	3.0855638

1. CD Account (Odds Ratio = 47.874649): Having a CD account has a substantial positive influence, with the odds of loan approval being nearly 48 times higher for those with a CD account.

Percentage Change: $(47.874649 - 1) \times 100 = 4687.46\%$

2. Education (Odds Ratio = 5.341812): Higher education levels greatly increase the odds of loan approval by more than five times.

Percentage Change: $(5.341812 - 1) \times 100 = 434.18\%$

3. Family (Odds Ratio = 1.995585): Each additional family member nearly doubles the odds of loan approval (99.56% increase).

Percentage Change: $(1.995585 - 1) \times 100 = 99.56\%$

4. Credit Card (Odds Ratio = 0.3240899): Having a credit card decreases the odds of loan approval.

Percentage Change: $(0.3240899 - 1) \times 100 = -67.59\%$

5. Securities Account (Odds Ratio = 0.4006267): Having a securities account would influence loan approval negatively.

Percentage Change: $(0.4006267 - 1) \times 100 = -59.94\%$

6. Online Banking (Odds Ratio = 0.52219): Using online banking decreases the odds of loan approval.

Percentage Change: $(0.52219 - 1) \times 100 = -47.78\%$

7. Income (Odds Ratio = 1.059078): For every unit increase in income, the odds of loan approval increase by approximately 5.91%.

Percentage Change: $(1.059078 - 1) \times 100 = 5.91\%$

Summary and Contrast of Rankings:

Comparing the two rankings from Statistical and Practical perspectives, they comply with each other and suggest the same ranks but different levels of influence for each factor.

While the statistically significant influence ranking emphasizes the impact of each predictor based on parameter estimates and Chi-square values, the practically significant influence ranking highlights the change in odds of loan approval associated with each predictor, expressed through odds ratios and percentage changes. Both perspectives agree on the relative importance of the predictors but differ in their interpretation of the direction and magnitude of influence.

For instance, the parameter estimates indicate that having a CD account (Estimate = -1.9342931) has a strong negative influence on loan approval, whereas the odds ratio (47.874649) suggests a substantial positive influence. This discrepancy may arise from the context in which the predictors are evaluated—parameter estimates focus on statistical significance while odds ratios provide a practical measure of effect size.

Similarly, higher education is highlighted as a critical factor in both rankings, with a positive estimate (1.67556497) and a significant increase in odds (Odds Ratio = 5.341812). This consistency reinforces the importance of education in loan approval decisions.

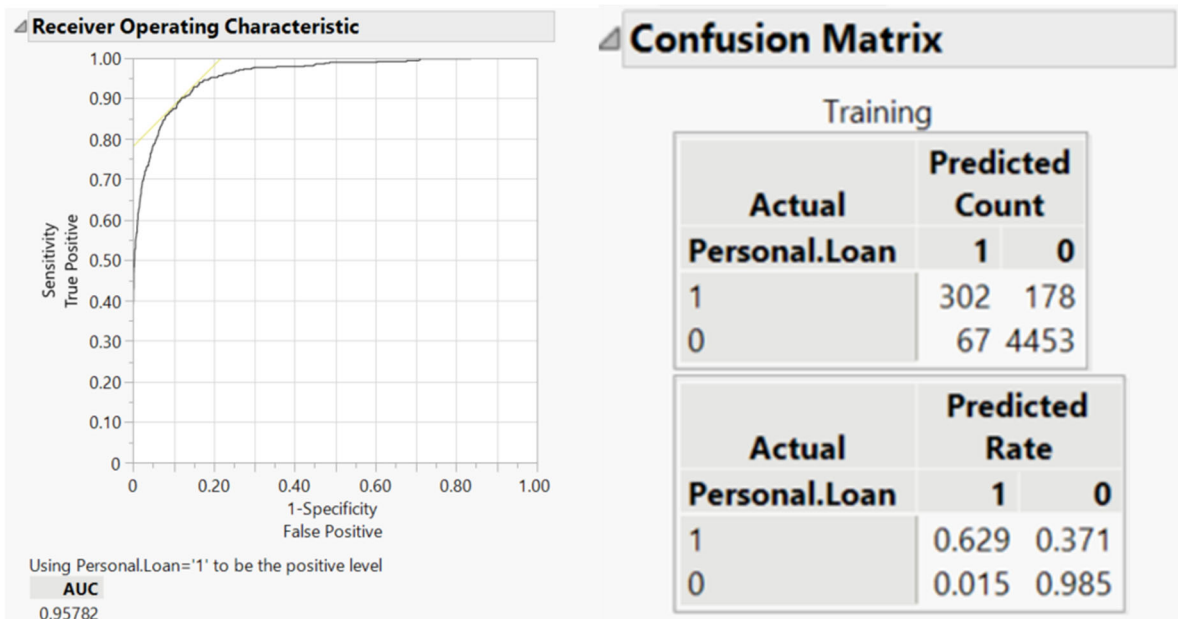
Family size also shows a strong positive influence in both rankings, with a positive estimate (0.69093737) and nearly doubled odds (Odds Ratio = 1.995585). This alignment suggests that larger families are consistently favored in loan approvals.

Conversely, the influence of having a credit card, a securities account, and using online banking shows some divergence. Statistically, these factors have positive parameter estimates, indicating a positive influence on loan approval. However, their odds ratios (0.3240899, 0.4006267, and 0.52219 respectively) suggest a negative impact on the odds of loan approval. These differences highlight the complexity of interpreting predictive influences and the importance of considering both statistical and practical perspectives.

Thus, while the statistical and practical significance rankings align in terms of the relative importance of predictors, they reveal different insights into the magnitude of these influences, underlining the necessity of a comprehensive analysis for accurate interpretation.

5. Discussion

5.1. Interpretation of Receiver Operating Characteristic (ROC) Curve and Confusion Matrix



Receiver Operating Characteristic (ROC) Curve:

ROC Curve: The ROC curve plots the True Positive Rate against the False Positive Rate for different threshold values. It is used to evaluate the performance of a classification model.

True Positive Rate: The proportion of actual positives that are correctly identified by the model (the percentage of correctly identified loan approvals).

False Positive Rate: The proportion of actual negatives that are incorrectly identified as positives by the model (the percentage of incorrectly identified loan approvals).

Performance: The ROC curve is close to the top-left corner, indicating high sensitivity and low false positive rate, which suggests a good model performance.

Area Under the Curve (AUC): The area under the ROC curve is a measure of the model's ability to distinguish between positive and negative classes. The Area Under the Curve (AUC) is 0.95782, which confirms that the model has a high ability to distinguish between approved and not approved loans, suggesting that the logistic regression model is highly effective in predicting loan approval status.

Confusion Matrix

Confusion Matrix: The confusion matrix shows the actual vs. predicted classifications for the training data, providing insight into the model's performance in terms of accuracy, precision, recall, and overall error rates.

True Positives (TP): The model correctly predicted 302 instances where a personal loan was approved (actual = 1, predicted = 1).

False Negatives (FN): The model incorrectly predicted 178 instances where a personal loan was not approved when it actually was (actual = 1, predicted = 0).

False Positives (FP): The model incorrectly predicted 67 instances where a personal loan was approved when it was not (actual = 0, predicted = 1).

True Negatives (TN): The model correctly predicted 4453 instances where a personal loan was not approved (actual = 0, predicted = 0).

Sensitivity: The model has a sensitivity of 0.629, meaning it correctly identifies 62.9% of the actual loan approvals.

Specificity: The model has a specificity of 0.985, meaning it correctly identifies 98.5% of the actual loan rejections.

Precision: The precision for loan approval is the proportion of predicted loan approvals that are actually approved, $TP / (TP + FP) = 302 / (302 + 67) = 0.818$.

Accuracy: The overall accuracy of the model is the proportion of correctly classified instances (both true positives and true negatives), $(TP + TN) / (TP + TN + FP + FN) = (302 + 4453) / (302 + 4453 + 67 + 178) = 0.951$.

Improvement: There is room for improvement in identifying loan approvals, as indicated by the false negatives (178 instances).

5.2. Implications for Risk Management

These findings align with the literature on big data's transformative potential in risk management. By leveraging extensive datasets and sophisticated models, financial institutions can enhance their risk management frameworks, reducing defaults and improving decision-making processes. Implementing these analytical methods can help institutions better assess the risk profiles of applicants, leading to more accurate and timely loan approval decisions.

6. Conclusion

The logistic regression analysis conducted in this study provides valuable insights into the factors that influence loan approval decisions by financial institutions. The findings highlight the significant role of income, family size, education, CD accounts, online banking, and credit card ownership in increasing the likelihood of loan approval. Specifically, applicants with higher income, more family members, higher education levels, and those actively engaged with financial products are more likely to have their loan applications approved. These results underscore the importance of a stable financial profile and active financial engagement in the loan approval process.

The study also reveals the nuanced impact of having a securities account, suggesting that while it is generally perceived as a positive indicator of financial stability, it may also signal the need for more liquid assets/money stream, thus affecting the loan approval decision. This nature emphasizes the complexity of financial behaviors and their interpretation by lenders.

By understanding these crucial predictors, financial institutions can refine their risk assessment models to make more accurate and fair loan approval decisions. The insights gained from this analysis can also inform applicants about the factors that significantly influence their chances of obtaining a loan. Future research could further explore the interplay between more financial behaviors and loan approval outcomes, incorporating more diverse datasets from different institutions to exclude the chance that the conclusion derived in this paper is a special case and other advanced analytical techniques in addition.

In conclusion, this study focuses on credit risk assessment and loan approval processes, providing actionable insights for both financial institutions and loan applicants. By recognizing and leveraging these critical factors, lenders can enhance their decision-making processes, and enable both borrowers and financial institutions to make decisions more efficiently.

Key Findings:

Logistic regression analysis effectively identifies key predictors that significantly influence loan approval decisions.

Higher income, larger family size, higher education levels, possession of a CD account, use of online banking, and having a credit card are strongly associated with an increased likelihood of loan approval.

Incorporating big data analytics enhances traditional risk management frameworks by incorporating diverse data sources and sophisticated analytical techniques to improve the accuracy and fairness of loan approval processes.

Future Research:

Future studies should explore integrating machine learning, decision tree, neural networks, and other techniques to further refine predictive models and address emerging challenges in credit risk management. Additionally, expanding the dataset to include more diverse borrower profiles could enhance the generalizability of the findings. Investigating the impact of macroeconomic factors on credit risk prediction could also provide valuable insights for financial institutions.

Practical Applications:

Financial institutions can apply these findings to develop more accurate and efficient risk assessment models, ultimately leading to better-informed lending decisions. By continuously refining their analytical approaches and incorporating new data sources, institutions can stay ahead of emerging risks and maintain competitiveness in the financial sector.

References:

- [1] Sabarostami, K. (2020). Risk Analytics in Banking. Kaggle. <https://www.kaggle.com/datasets/sabarostami/risk-analytics-in-banking>
- [2] Hasnat, B. (2018). Big data: An institutional perspective on opportunities and challenges. *Journal of Economic Issues*, 52(2), 580-588.
- [3] Sagioglu, S., & Sinanc, D. (2013). Big data: A review. San Diego, CA, USA: International Conference on Collaboration Technologies and Systems.
- [4] Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. (2015). Big data in accounting: An overview. *Accounting Horizons*, 29, 381-396.
- [5] World Bank. (2020). Global Financial Inclusion Database.
- [6] Chen, J., Tao, Y., Wang, H., & Chen, T. (2015). Big data based fraud risk management at Alibaba. *The Journal of Finance and Data Science*, 1, 1-10.
- [7] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.