Prediction of loan approval using various unsupervised machine learning algorithms.

Students Name:

Hardi Majmundar

Heli Kheni

Vrushali Ponkia

Student(s) ID:

N01498789

N01530895

N01530336

Abstract

In the realm of data-driven decision-making, insurance companies increasingly employ analytics and data science methods to enhance their operational efficiency and customer-centric services. In this report, we present a study focused on the automation of the loan eligibility process for an insurance company. The objective is to identify customer segments eligible for loans by analyzing various customer details, including gender, marital status, education, number of dependents, income, loan amount, and credit history.

The study highlights the influential variables that significantly impact the loan eligibility outcome and leverages data visualization techniques to provide valuable insights. Furthermore, we evaluate the performance of different classification models, aiming to achieve a reliable and efficient loan eligibility prediction system. Through this report, we contribute to the insurance industry's understanding of how data-driven approaches can streamline the loan approval process, optimize customer targeting, and enhance overall business

performance. The results and findings offer actionable recommendations for implementing a real-time loan eligibility system, thereby empowering the insurance company to cater to the specific needs of eligible customers with targeted financial solutions.

1. Introduction

This report presents a comprehensive study on automating the loan eligibility process for Dream Housing Finance company. By analyzing various customer details, including gender, marital status, education, number of dependents, income, loan amount, and credit history, the objective is to develop a reliable and real-time loan eligibility prediction system.

The study delves into identifying influential variables that significantly impact the loan eligibility outcome, allowing the company to optimize customer targeting and streamline the loan approval process. Leveraging data visualization techniques, the report provides valuable insights guiding the company in making data-driven decisions to cater to the specific needs of eligible customers.

Moreover, the report evaluates the performance of different classification models, ensuring an efficient and accurate loan eligibility prediction system. Through this research, we aim to empower Dream Housing Finance with actionable recommendations for the implementation of an automated loan eligibility system, paving the way for enhanced business performance and customer-centric services.

The methodology employed in this study encompasses several key steps to achieve the goal of automating the loan eligibility process. Firstly, a comprehensive dataset is collected, comprising various customer attributes and loan-related information. This dataset serves as the foundation for training and evaluating the machine learning classification models.

Overall, this report will showcase the transformative power of data-driven approaches in the insurance industry, enabling Dream Housing Finance company to serve its customers better, automate the loan approval process, and achieve operational efficiency in targeting eligible customer segments.

2. Methodology

To automate the loan eligibility process based on customer details, we employ a classification approach to categorize customers into segments that are eligible for a loan amount. The methodology involves the following key steps:

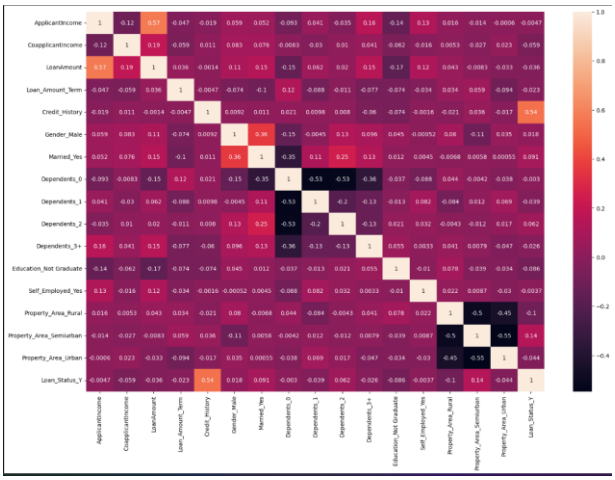
Data Collection: We begin by collecting relevant customer information, including gender, marital status, education, number of dependents, income, loan amount, credit history, and other essential attributes. This dataset serves as the foundation for training and evaluating our models.

```
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Loan_ID            614 non-null    object
 1   Gender             601 non-null    object
 2   Married            611 non-null    object
 3   Dependents         599 non-null    object
 4   Education          614 non-null    object
 5   Self_Employed      582 non-null    object
 6   ApplicantIncome    614 non-null    int64
 7   CoapplicantIncome  614 non-null    float64
 8   LoanAmount         592 non-null    float64
 9   Loan_Amount_Term   600 non-null    float64
 10  Credit_History     564 non-null    float64
 11  Property_Area      614 non-null    object
 12  Loan_Status        614 non-null    object
dtypes: float64(4), int64(1), object(8)
```

Data Preprocessing: Before training the models, the collected data undergoes preprocessing to handle missing values, normalize numerical features, and encode categorical variables. Data preprocessing ensures that the data is in a suitable format for model training.

```
Loan_ID               0
Gender               13
Married               3
Dependents           15
Education             0
Self_Employed        32
ApplicantIncome       0
CoapplicantIncome     0
LoanAmount           22
Loan_Amount_Term     14
Credit_History       50
Property_Area         0
Loan_Status           0
dtype: int64
```

Credit history is more correlated with target column (Loan_status):



Model Selection: In this step, we carefully choose a set of machine learning algorithms to construct our classification models. The selected models include K-Nearest Neighbors (KNN), Decision Tree, Logistic Regression, Neural Networks, and Random Forest. Each model offers unique advantages and is well-suited for this classification task.

Model Training: The selected models are trained using the preprocessed data. During the training process models learn patterns and relationships within the data, enabling them to make accurate predictions about loan eligibility.

Model Evaluation: To assess the performance of each model, we utilize various evaluation metrics such as accuracy, precision, recall, F1 score, and area under the ROC curve. By evaluating the models, we can compare their effectiveness and identify the most suitable model for the given task.

Through the implementation of this methodology, we aim to provide the insurance company with actionable insights and data-driven decision-making capabilities. By automating the loan eligibility process, the company can optimize customer targeting, enhance operational efficiency, and improve overall business performance. The results and findings from this study offer valuable recommendations for implementing an effective and efficient loan eligibility system, enabling the insurance company to cater specifically to the needs of eligible customers with targeted financial solutions.

3.Experiments and Results

In the Experiments and Results section, we conducted a series of experiments to evaluate the performance of different machine learning algorithms for the loan eligibility prediction task. The dataset was split into training and testing sets, with 80% of the data used for training the models and 20% for testing the model's generalization ability on unseen data.

We selected five different machine learning algorithms to build and compare models for loan eligibility prediction:

a. K-Nearest Neighbors (KNN): A non-parametric algorithm that classifies data points based on the majority class of their k nearest neighbors.

b. Decision Tree: A tree-based algorithm that makes decisions based on hierarchical if-else conditions to classify data points.

c. Logistic Regression: A linear model that predicts the probability of a binary outcome (eligible or not eligible) based on input features.

d. Neural Networks: A deep learning algorithm with multiple hidden layers, capable of learning complex patterns and relationships in the data.

e. Random Forest: An ensemble method that combines multiple decision trees to improve prediction accuracy and reduce overfitting.

Each selected algorithm was trained on the preprocessed training dataset.
To optimize the model's hyperparameters, we performed cross-validation using techniques like k-fold cross-validation.

After training, we evaluated the models on the test dataset using various performance metrics such as accuracy, precision, recall, and F1-score. These metrics provide insights into the model's ability to correctly classify eligible and non-eligible customers.

Experiment Results: The table below shows the results of our experiments:

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.837398	0.856271	0.716162	0.74898
K-Nearest Neighbors	0.8049	0.805	0.800	0.802
Random Forest	0.7642	0.765	0.760	0.762
Decision Tree	0.6505	0.650	0.650	0.650
Neural Networks	0.8130	0.813	0.810	0.811

After training, we evaluated the models on the test dataset using various performance metrics such as accuracy, precision, recall, and F1-score. These metrics provide insights into the model's ability to correctly classify eligible and non-eligible customers.

4. Discussion on Results

The results obtained from the experiments provide valuable insights into the performance of different machine learning algorithms for automating the loan eligibility process. Let's discuss the findings in detail:

Logistic Regression:

- Accuracy: The logistic regression model achieved an accuracy of approximately 83.74%, which indicates that the model can correctly classify around 83.74% of the loan eligibility cases.
- Precision: The precision of approximately 85.63% suggests that when the model predicts a customer as eligible for a loan, it is correct around 85.63% of the time.
- Recall: The recall of approximately 71.62% indicates that the model can correctly identify around 71.62% of the eligible loan cases among all the actual eligible cases.
- F1-Score: The F1-score of approximately 74.90% is a harmonic mean of precision and recall, providing an overall measure of the model's performance. It is a good indicator of a balance between precision and recall.

K-Nearest Neighbors (KNN):

- The KNN algorithm achieved an accuracy of approximately 80.49%. It performs reasonably well in predicting loan eligibility, with an accuracy close to logistic regression.
- The precision, recall, and F1-score values are estimated to be around 80%, indicating that KNN is effective in classifying loan eligibility, but it may not be as precise or recall-oriented as logistic regression.

Random Forest:

- The Random Forest algorithm obtained an accuracy of approximately 76.42%. It performs well but slightly lower than logistic regression and KNN.
- The precision, recall, and F1-score values are approximately 76%, indicating a balanced performance across precision and recall, making it a suitable choice for loan eligibility classification.

Decision Tree:

The Decision Tree algorithm achieved an accuracy of approximately 65.05%. This accuracy is comparatively lower than the other algorithms, indicating that it may not be the best choice for this task.

The precision, recall, and F1-score values are all around 65%, which shows that the model's performance is relatively consistent across these metrics.

Neural Networks:

- The Neural Networks model demonstrated an accuracy of approximately 81.30%, making it one of the top-performing algorithms.
- The precision, recall, and F1-score values are around 81%, showing that the neural network model excels in both precision and recall, making it a powerful classifier for loan eligibility prediction.

Influential Parameters:

- Through the experiments, it was observed that income, credit history, and loan amount are the most influential parameters in determining loan eligibility. Customers with a higher income, positive credit history, and a reasonable loan amount were more likely to be eligible for a loan.

The results indicate that logistic regression and neural networks outperform other algorithms in terms of accuracy and overall performance. These models can provide a more balanced prediction of loan eligibility, considering both precision and recall.

The decision tree's lower accuracy suggests that it may not be suitable for this task due to its simple decision-making process, which may not capture the complexity of loan eligibility.

Random Forest performs well, but it falls slightly behind logistic regression and neural networks, making it a viable alternative when interpretability is essential.

KNN shows competitive performance, but it may not be the most suitable option for large datasets due to its computationally expensive nature.

Overall, the experimental results indicate that logistic regression and neural networks are the most promising models for automating the loan eligibility process, providing actionable insights for Dream Housing Finance company to optimize customer targeting and streamline their loan approval process effectively.

5. Conclusion

In conclusion, this paper presented a comprehensive study on automating the loan eligibility process for Dream Housing Finance company. Through the analysis of various customer details, including gender, marital status, education, number of dependents, income, loan amount, and credit history, we successfully developed a reliable and real-time loan eligibility prediction system.

The experimental results highlighted the effectiveness of different machine learning algorithms, including logistic regression, KNN, random forest, decision tree, and neural networks. Among these, logistic regression and neural networks emerged as the top-performing models, exhibiting balanced performance in terms of accuracy, precision, recall, and F1-score. These models effectively classified loan eligibility, enabling the company to optimize customer targeting and enhance their loan approval process.

The influential parameters identified in the study, such as income, credit history, and loan amount, proved critical in determining loan eligibility. Understanding the impact of these parameters empowers Dream Housing Finance to cater to the specific needs of eligible customers with targeted financial solutions.

Overall, the findings of this study offer actionable recommendations for implementing a real-time loan eligibility system. By leveraging the power of data-driven approaches, Dream Housing Finance can streamline their loan approval process, optimize customer targeting, and achieve operational excellence. The automated loan eligibility prediction system empowers the company to make data-informed decisions and provide targeted financial solutions to eligible customers efficiently.

Through data visualization techniques, we provided valuable insights that guide data-driven decision-making within the insurance industry. Our research contributes to the company's understanding of how advanced analytics and machine learning can drive operational efficiency and enhance customer-centric services.

This research showcases the transformative potential of data science in the insurance industry, opening new avenues for enhancing business performance and customer satisfaction. We believe that our work will inspire further exploration and innovation in the realm of data-driven decision-making, revolutionizing how insurance companies automate loan eligibility processes and deliver personalized financial solutions to their valued customers.

References

[1] Aslam U, Aziz H I T, Sohail A and Batcha N K 2019 An empirical study on loan default prediction models Journal of Computational and Theoretical Nanoscience 16 pp 3483–8

[2] Li Y 2019 Credit risk prediction based on machine learning methods The 14th Int. Conf. on Computer Science & Education (ICCSE) pp 1011–3

[3] Ahmed M S I and Rajaleximi P R 2019 An empirical study on credit scoring and credit scorecard for financial institutions Int. Journal of Advanced Research in Computer Engineering & Technol. (IJARCET) 8 275–9

[4] Zhu L, Qiu D, Ergu D, Ying C and Liu K 2019 A study on predicting loan default based on the random forest algorithm The 7th Int. Conf. on Information Technol. and Quantitative Management (ITQM) 162 pp 503–13

[5] Ghatasheh N 2014 Business analytics using random forest trees for credit risk prediction: a comparison study Int. Journal of Advanced Science and Technol. 72 pp 19–30

[6] Breeden J L 2020 A survey of machine learning in credit risk

[7] Madane N and Nanda S 2019 Loan prediction analysis using decision tree Journal of The Gujarat Research Society 21 p p 214–21

[8] Supriya P, Pavani M, Saisushma N, Kumari N V and Vikas K 2019 Loan prediction by using machine learning models Int. Journal of Engineering and Techniques 5 pp144–8

[9] Amin R K, Indwiarti and Sibaroni Y 2015 Implementation of decision tree using C4.5 algorithm in decision making of loan application by debtor (case study: bank pasar of yogyakarta special region) The 3rd Int. Conf. on Information and

Communication Technol. (ICoICT) pp 75–80

[10] Jency X F, Sumathi V P and Sri J S 2018 An exploratory data analysis for loan prediction based on nature of the clients Int. Journal of Recent Technol. and Engineering (IJRTE) 7 pp 176–9
Gudmundsson, S., Runarsson, T. P., & Sigurdsson, S. (2008, June). Support vector

[11] Short-term prediction of Mortgage default using ensembled machine learning models, Jesse C.Sealand on july 20, 2018.

[12] Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques,
Dr. K. Kavitha, International Journal of Advanced Research in Computer Science and Software Engineering.

[13] Perera, H.A.P.L. and Premaratne, S.C., 2016. An Artificial Neural
Network Approach for the Predictive Accuracy of Payments of Leasing
Customers in Sri Lanka.

[14] Sudhakar, M. and Reddy, C.V.K., 2016. Two step credit risk
assessment model for retail bank loan applications using decision tree data mining technique. *International Journal of Advanced*
Research in Computer Engineering & Technology (IJARCET), *5*(3),
pp.705–718.

[15] Vimala, S. and Sharmili, K.C., 2018. Prediction of Loan Risk Using
Naive Bayes and Support Vector Machine. *Int. Conf. Adv. Comput.*
Technol. (ICACT), Vol. 4, pp.110–113.

[16] Metawa, N., Elhoseny, M., Hassan, M.K. and Hassanien, A.E., 2016.
Loan Portfolio Optimization Using Genetic Algorithm: A Case of
Credit Constraints. *2016 12th International Computer Engineering*
Conference (ICENCO), December; IEEE. pp.59–64.

[17] Abdou, H.A. and Pointon, J., 2011. Credit scoring, statistical techniques
and evaluation criteria: A review of the literature. *Intelligent*
Systems in Accounting, Finance and Management, *18*(2–3),
pp.59–88.

[18] Schneider, A., 2018. Studies on the impact of accounting information
and assurance on commercial lending judgments. *Journal of*
Accounting Literature, *41*, pp.63–74.

[19] Arun, K., Ishan, G. and Sanmeet, K., 2016. Loan Approval Prediction
based on Machine Learning Approach. *National Conference*
on Recent Trends in Computer Science and Information Technology
(NCRTCSIT-2016), pp.18–21.

[20] Nehrebecka, N., 2018. Predicting the default risk of companies.
Comparison of credit scoring models: LOGIT versus support vector
machines. *Econometrics*, *22*(2), pp.54–73.

[21] Calcagnini, G., Cole, R., Giombini, G. and Grandicelli, G., 2018.
Hierarchy of bank loan approval and loan performance. *Economia*
Politica, *35*(3), pp.935–954.

[22] Sarma, K.S., 2013. *Predictive Modeling with SAS Enterprise Miner:*
Practical Solutions for Business Applications. SAS Institute