Predictive Analytics for Financial Services

Team – Nu

ABSTRACT

The recent expansion of the credit industry has made credit scoring a very important problem, so the bank's credit section deals with a lot of credit data. There are numerous data mining techniques that have been suggested so far to address credit scoring issues, and each of them has some advantages and disadvantages over the others. However, there isn't a comprehensive reference outlining the most popular data mining techniques for credit scoring issues. This project aims to develop a model for classifying bank customers as "good" or "bad" based on their credit scores and other relevant details. The dataset used for this project contains information on the customers' credit scores, income, debt-to-income ratio, employment status, and other factors that may affect their creditworthiness. The project involves pre-processing the data, performing feature selection and engineering, and training various machine learning models, including a Dummy Classifier, KNN, decision trees, SVC and random forests. The models' performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. The results of the project can help banks improve their risk assessment process, identify potential defaulters, and make more informed lending decisions.

Index Terms—Classification, data mining techniques, predictive analytics, credit scoring

INTRODUCTION

Credit risk assessment is an essential task for banks to make informed lending decisions and minimize the risk of losses. However, with the recent expansion of the credit industry, banks deal with a large volume of credit data that cannot be analysed manually. Therefore, data mining methods have been proposed as an effective solution for credit scoring issues. However, there is still a need for a comprehensive reference outlining the most popular data mining techniques for credit risk assessment. Accurate credit risk assessment is critical for banks to identify potential defaulters and make more informed lending decisions. By doing so, banks can improve their risk assessment process and make better use of their resources. Therefore, it is important to study and solve this problem to benefit the banking sector as a whole. In this project, we aim to develop a model for classifying bank customers as "good" or "bad" based on their credit scores and other relevant details. To achieve this, we will pre-process the data, perform feature selection and engineering, and train various machine learning models, including a Dummy Classifier, KNN, decision trees, SVC, and random forests. The models' performance will be evaluated using metrics such as accuracy, precision, recall, and F1-score. The dataset used for this project contains information on the customers' credit scores, income, debt-to-income ratio, employment status, and other factors that may affect their creditworthiness. We will pre-process the data by dealing with missing values, handling categorical variables, and scaling numerical features. Feature selection and engineering will be performed to identify the most relevant features and transform them to improve model performance. We will train various machine learning models and compare their performance to identify the best model for credit risk assessment. The models' performance will be evaluated using metrics such as accuracy, precision, recall, and F1-score. The results of this project can help banks improve their risk assessment process, identify potential defaulters, and make more informed lending decisions. By providing a comprehensive reference for data mining techniques used in credit scoring, this project will contribute to the development of more effective credit risk assessment methods in the future. It will help the banking sector to make better use of data mining techniques and improve their risk assessment process. This, in turn, will benefit the economy by reducing the risk of losses for banks and increasing the availability of credit for businesses and individuals. In conclusion, credit risk assessment is an important problem domain in the banking sector that requires accurate and efficient solutions. In this project, we aim to develop a model for classifying bank customers as "good" or "bad" based on their credit scores and other relevant details. The results of this project can help banks improve their risk assessment process, identify potential defaulters, and make more informed lending decisions. By providing a comprehensive reference for data mining techniques used in credit scoring, this project will contribute to the development of more effective credit risk assessment methods in the future.

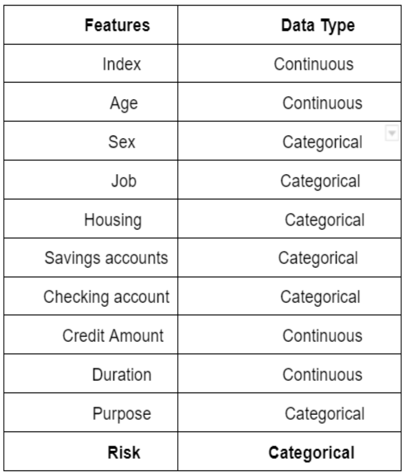
RELATED WORK

[1]Broby (2022) discussed the use of predictive analytics in finance, specifically in credit risk assessment. The author highlighted the importance of accurate credit risk assessment in making informed lending decisions and minimizing the risk of losses for financial institutions. The article also discussed various data mining techniques used in credit scoring, such as logistic regression,

decision trees, and neural networks. In addition, Broby (2022) noted the importance of feature selection and engineering in improving the performance of credit scoring models. The author emphasized the need for using domain knowledge and expert judgment in selecting relevant features for credit risk assessment. Overall, Broby's (2022) article provides valuable insights into the use of predictive analytics in finance, specifically in credit risk assessment. The article highlights the importance of accurate credit risk assessment and discusses various data mining techniques and feature selection methods used in this domain. This article will serve as a useful reference for our project in developing a decision tree model for classifying bank customers as "good" or "bad" based on their credit scores and other relevant details. [2]The paper titled "A Proposed Classification of Data Mining Techniques in Credit Scoring" proposes a classification of data mining techniques used in credit scoring. The authors of this paper categorized the data mining techniques into two groups: statistical and machine learning methods. The statistical methods include logistic regression, discriminant analysis, and Bayesian networks, while the machine learning methods include decision trees, neural networks, and support vector machines. The paper also discussed the advantages and limitations of each data mining technique and highlighted the need for selecting the appropriate method based on the characteristics of the data and the goals of the analysis. In addition, the authors emphasized the importance of feature selection and engineering in credit scoring and discussed various techniques used for this purpose. The proposed classification of data mining techniques will help us in selecting the appropriate method for our analysis, and the discussion on feature selection and engineering will guide us in improving the performance of our model. [3] In relation to the current project, the study by Indriasari et al. (2019) provides valuable insights into the use of machine learning techniques for credit risk assessment in the banking sector. The authors' emphasis on feature selection and engineering aligns with the approach taken in the current project. The study also highlights the importance of evaluating the performance of machine learning models using appropriate metrics, which is a crucial aspect of the current project. Overall, this study provides a useful reference for understanding the potential of predictive analytics in finance and its applications in credit risk assessment.

DATA

The data set is the “German credit risk data” which is available in Kaggle repository. This dataset consists of 9 attributes and 1000 instances. In this dataset, each entry represents a person who takes a credit by a bank. Each person is classified as good or bad credit risks according to the set of attributes.



Features	**Data Type**
Index	Continuous
Age	Continuous
Sex	Categorical
Job	Categorical
Housing	Categorical
Savings accounts	Categorical
Checking account	Categorical
Credit Amount	Continuous
Duration	Continuous
Purpose	Categorical
Risk	**Categorical**

Table:1

Only two of the attributes had NA values. The NA values were dropped from columns checking account and saving account for accurate classification of customers. The following are the exploratory data analysis of the categorical attributes present in the data set. From the sex bar graph, we could interpret that the probability of a male being a good customer is higher than that of female. Looking at the various jobs’ customers have; the skilled professionals are most probably good customers. It is highly possible that a customer who owns a house and has little to moderate amount in the savings account are ‘good’ customers. Fig:5 The correlation heat map is extracted after the data pre-processing and shows the attributes which are highly correlated to other. It also shows that the target variable is the attribute ‘RISK’.

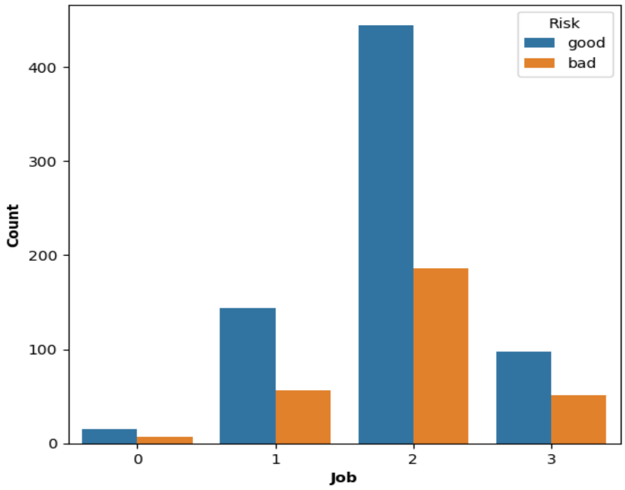


Fig:1

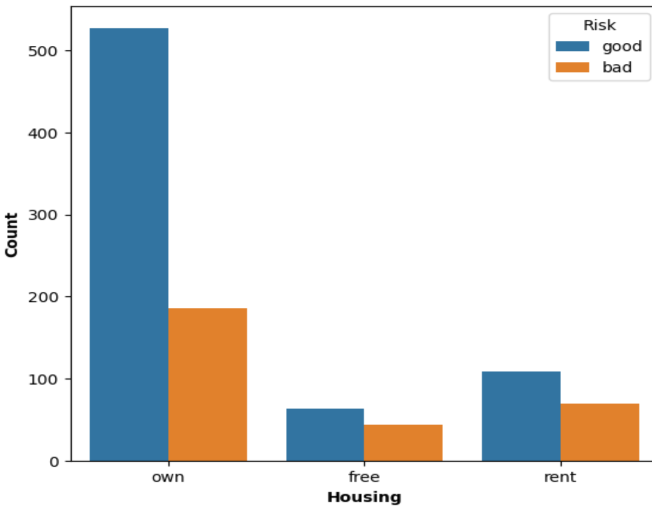


Fig:2

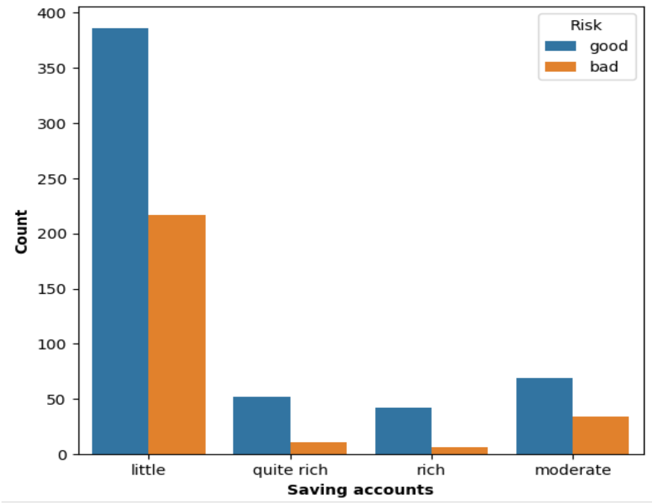


Fig:3

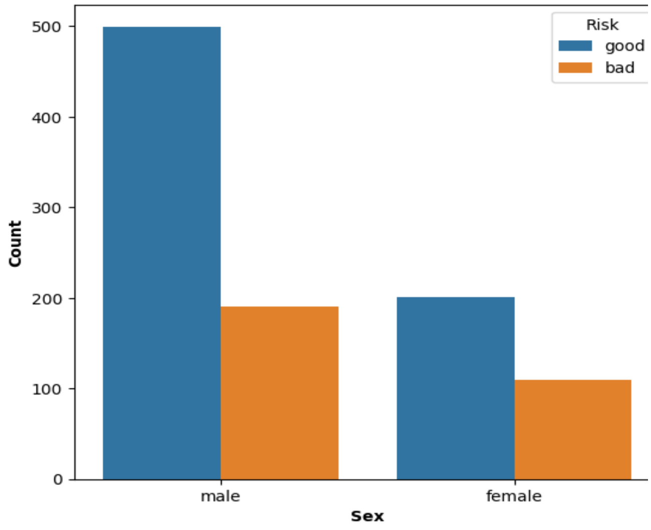


Fig:4

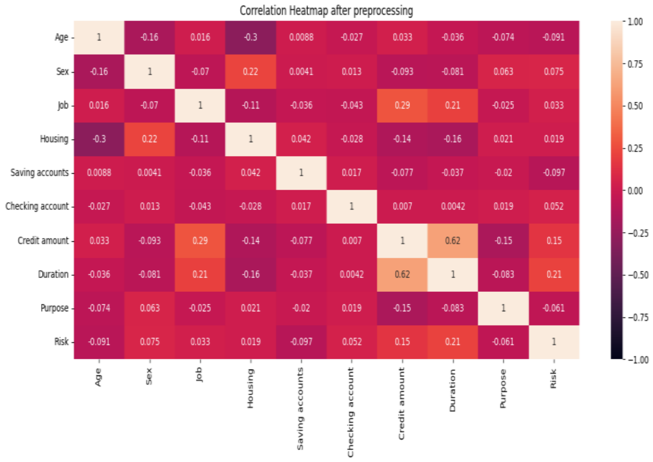


Fig:5

The following are the visualizations of quantitative attributes like age, credit amount, duration. We have observed few outliers which were dealt by replacing them with the highest value of the particular attribute.

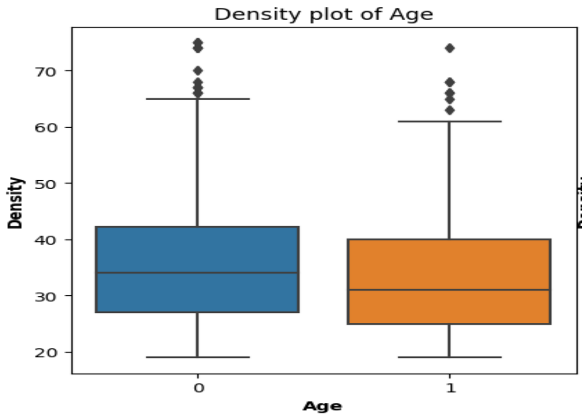


Fig:6

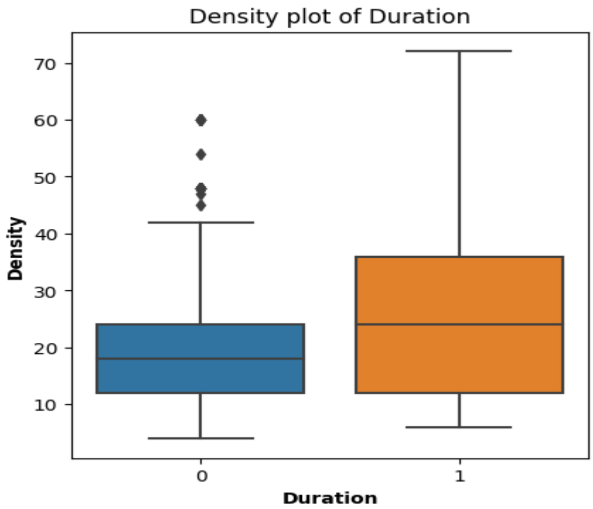


Fig:7

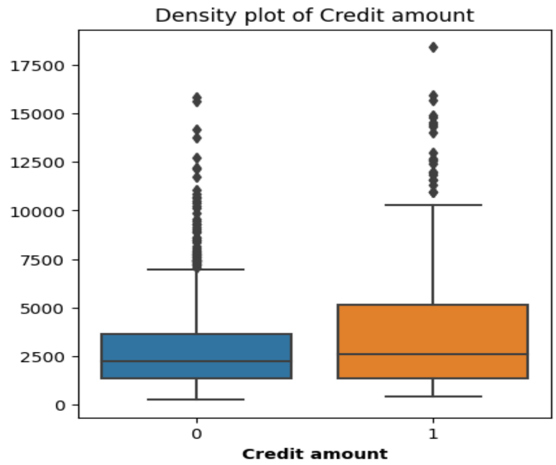


Fig: 8

METHODS

In this section, we emphasized on the pre-processing of data by using one-hot encoding, also tried to use SMOTE to deal with imbalanced data.

- One-hot encoding

We have used one-hot encoding, which is a method used in data analysis and machine learning to encode categorical variables as numerical data. Each categorical value in one-hot encoding is transformed into a binary vector of 1s and 0s, with the length of the vector being equal to the number of distinct categories in the variable.

One-hot encoding's key benefit is that by portraying categorical data as numerical data, it enables machine learning algorithms to learn from that data. Additionally, it prevents the development of an ordinal relationship between categories, which might not be appropriate for datasets.

- SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is a well-liked data preprocessing method for imbalanced classification issues. When there are significantly less examples in one class than there are in another, this is known as an unbalanced categorization. The fundamental concept of SMOTE is to interpolate between pairs of existing samples from the minority class to construct synthetic samples.

In our case, we have a greater number of good customers than bad customers, this will bias the results. So, to avoid this condition we have implemented SMOTE technique to balance the imbalanced data.

We have tried and tested few classification models like

1)Random Forest:

A machine learning approach called a random forest classification model employs numerous decision trees to produce predictions. Each decision tree in this model is trained on a random subset of the data and features available, which helps to reduce overfitting and boost prediction accuracy.

2)K-Nearest Neighbors:

The supervised learning algorithm K-nearest neighbors (KNN) is a type used for classification and regression. By comparing a given observation to previously observed data points, classification attempts to predict the class to which it belongs.

3)SVM:

An effective supervised machine learning approach for classification and regression analysis is the Support Vector Machine (SVM). In a high-dimensional feature space, SVM aims to locate a hyperplane that maximum separates two classes.

4)Decision Tree:

A supervised machine learning approach called decision trees is utilized for classification and regression tasks. They describe a set of guidelines used to forecast the value of a target variable depending on several input features graphically.

EXPERIMENTS AND RESULTS

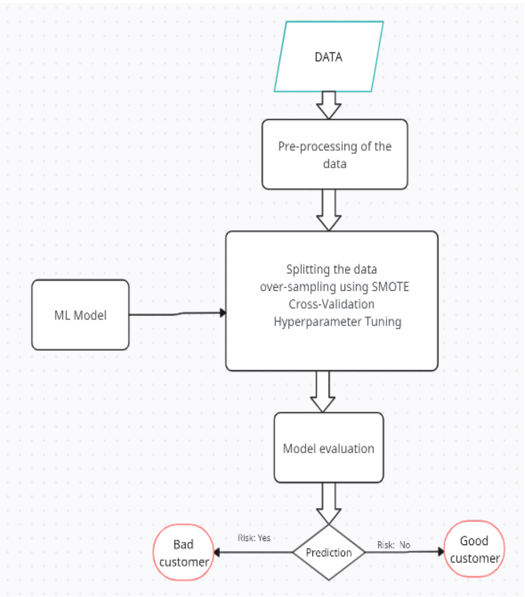


Fig: 9

Fig: 9 shows the experimental design of the project. After the data set is extracted from Kaggle, we have pre-processed the data using the techniques mentioned in the previous section. The data was split into training and testing set where 70% of the data was used for training and 30% for testing. Standard scaler was used for both original data and SMOTE data to normalize the data, improve the model's performance. Before we start training and testing other models, we have tried to use a dummy classifier to create a baseline for other performance and also to be sure if the other classification models were learning from the training set. We have used GridSearchCV to tune and select the best hyperparameters for each model. The following table shows the best hyperparameters for Random Forest, SVM, Decision Tree, KNN.

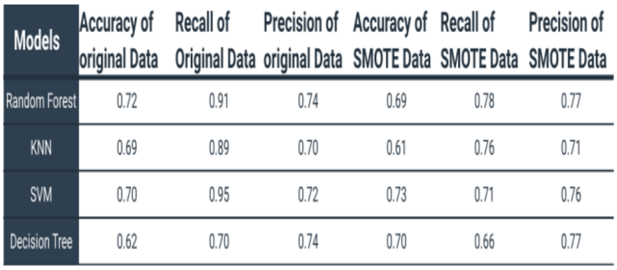
KNN	Decision Tree
Best Hyperparameters: {'n_neighbors': 1}	'criterion': 'gini' 'max_depth': 20
Best Accuracy Score: 0.686734693877551	'min_samples_leaf': 1, 'min_samples_split': 2

Table: 2

SVM	Random Forest
'C': 10	'max_depth':None 'max_features': 'auto' 'min_samples_leaf': 1
'Kernel': 'rbf'	'min_samples_split': 4 'n_estimators': 162

Table: 3

Accuracy, recall, precision, F1-score were used for evaluating the classification models. We have considered all the metrics generated by different models for both original data and the SMOTE data. The following table compares the evaluation metrics for original and SMOTE data.



Models	Accuracy of original Data	Recall of Original Data	Precision of original Data	Accuracy of SMOTE Data	Recall of SMOTE Data	Precision of SMOTE Data
Random Forest	0.72	0.91	0.74	0.69	0.78	0.77
KNN	0.69	0.89	0.70	0.61	0.76	0.71
SVM	0.70	0.95	0.72	0.73	0.71	0.76
Decision Tree	0.62	0.70	0.74	0.70	0.66	0.77

Table: 4

From the table we can interpret that the SMOTE data had comparatively lesser values accuracy, recall, precision because when SMOTE is performed it actually can generate noisy samples that do not accurately represent the minority class. This can lead to misclassification and degrade the performance of the model. Another reason can also be that the models like KNN, Decision Tree are sensitive to SMOTE technique.

Fig: 10 shows the comparison of the four models against the accuracy, precision and recall values. Of all the metrics we chose the accuracy score as the final metric to select the best model.

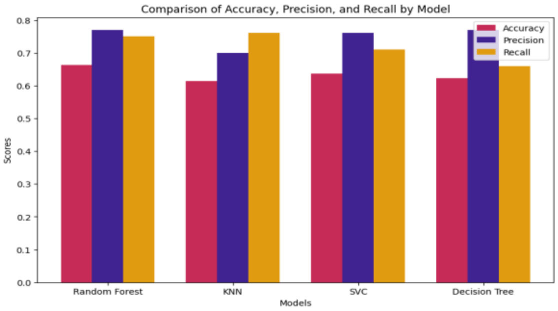


Fig:10

Fig: 11 provides a better understanding on, which model performs the best. If closely noticed, we can observe that the Random Forest has performed better than the dummy classifier and other models as well with an accuracy score of 0.72. Where as the accuracy of the dummy classifier is just 0.70 which explains that the model is learning.

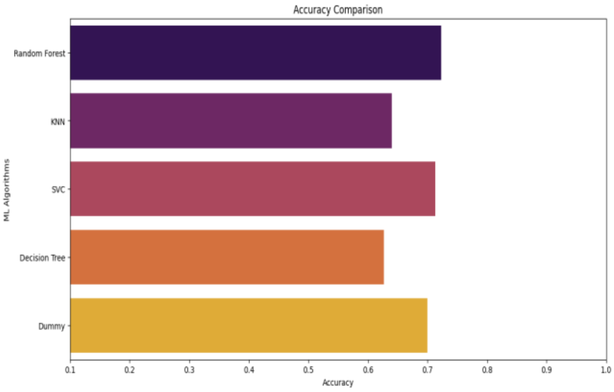


Fig: 11

CONCLUSION

After considering the evaluation metrics for SMOTE data and the original data, we have come to a conclusion that the Random Forest is the best model to predict if the customer is 'good' or 'bad' based on the accuracy scores. We also interpreted that the SMOTE data did not perform well in terms of accuracy, recall and precision. While the one-hot encoding has made it easy to normalize data to numerical values. In future, we would try and test other classification models like AdaBoost, Gradient Boost, XGBoost in order to improve the accuracy of the prediction.

REFERENCES

[1] Broby, D. (2022). The use of predictive analytics in Finance. The Journal of Finance and Data Science, 8, 145– 161. https://doi.org/10.1016/j.jfds.2022.0 5.003

[2] A proposed classification of data mining techniques in credit scoring. (n.d.). Retrieved April 1, 2023, from https://www.researchgate.net/publication/268049977_A_Proposed_Classification_of_Data_Mining_Techniques_in_Credit_Scoring

[3] E. Indriasari, H. Soeparno, F. L. Gaol and T. Matsuo, "Application of Predictive Analytics at Financial Institutions: A Systematic Literature Review," 2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI), Toyama, Japan, 2019, pp. 877-883, doi: 10.1109/IIAI-AAI.2019.00178