Credit

by H M

Submission date: 10-Jan-2023 10:47AM (UTC-0600)
Submission ID: 1990798730
File name: Credit.docx (21.28K)
Word count: 2314
Character count: 12310

UNIVERSITY OF SOUTHAMPTON

Credit Risk & Data Analytics

Student ID

Date YYYY

UNIT LEADER

Introduction

Credit scoring is an important tool for lenders to assess the risk of borrowers and identify the likelihood of them defaulting on their debts. It involves the analysis of a variety of factors, such as credit history, income, employment status and other financial indicators, to determine the creditworthiness of an individual. Credit scoring models help lenders in making more informed decisions, while also providing a more efficient and cost-effective way to manage their risk exposure.

The purpose of this report is to build an intuitive and predictive scorecard using a logistic regression classifier. The dataset used for this report is ‘Credit data.xlsx’ which contains data on 10,000 borrowers and whether they subsequently experienced serious delinquency (see variable ‘SeriousDlqin2yrs’). The report will cover the pre-processing of the dataset, building a scorecard using a logistic regression classifier and compare this scorecard with the result of a Random Forest model run over the data.

Question 2 (20mks)

Data Preprocessing

Exploratory data analysis is the first step of pre-processing. It involves understanding the data, its characteristics and relationships between different variables. This can be done by visualizing the data using various methods such as histograms, box plots, scatter plots, etc. This helps in understanding the data better and identifying any anomalies or patterns that may be present in the data.

The next step is to check for missing values in the data. This can be done by using summary statistics, such as mean, median, mode, etc. of each variable. If the summary statistics are not available for a particular variable, then it is likely that the data is missing for that variable. In this case, suitable methods for handling missing values need to be used.

Outliers can have a significant impact on the performance of the model. Therefore, it is important to detect and treat any outliers that may be present in the data. This can be done by using statistical methods such as the box plot and the interquartile range (IQR). Any data points that lie outside of the IQR can be considered as outliers and treated accordingly.

Binning the variables is a technique used to group data into bins or ranges. This can be done by grouping the data into meaningful intervals, such as age groups, income ranges, etc. It helps in simplifying the data and making it easier to model.

Coding the discrete variables using Weights of Evidence is a method used to convert discrete variables into numerical values. This is done by assigning weights to each discrete variable based on the probability of a particular outcome.

Finally, the dataset needs to be split into a training and test set. This can be done by using a technique such as cross-validation or simple random sampling. The training dataset is used to train the model and the test dataset is used to evaluate its performance.

Missing value handling method

The missing value handling method used in this case is imputing the missing values using the median of the respective columns. This method is advantageous when the missing values are very few and not distributed randomly. It brings the data into a usable form and is faster compared to other methods like dropping the rows/columns or using a machine learning model to predict the missing values.

The median is a robust measure of the central tendency, and it is performed on data points that are not affected by outliers. This helps in getting a more accurate value to fill the missing values. Also, the median can be used even when the distribution of the data points is not normal. It is easy to compute and understand, and requires little computational power.

In this case, we used the median for imputing the missing values in the 'MonthlyIncome' and 'NumberOfDependents' columns. This method is suitable as it helps to preserve the integrity of the data set and is faster than other methods. It is also less prone to errors, as it does not use any complex algorithms.

Outlier Treatment

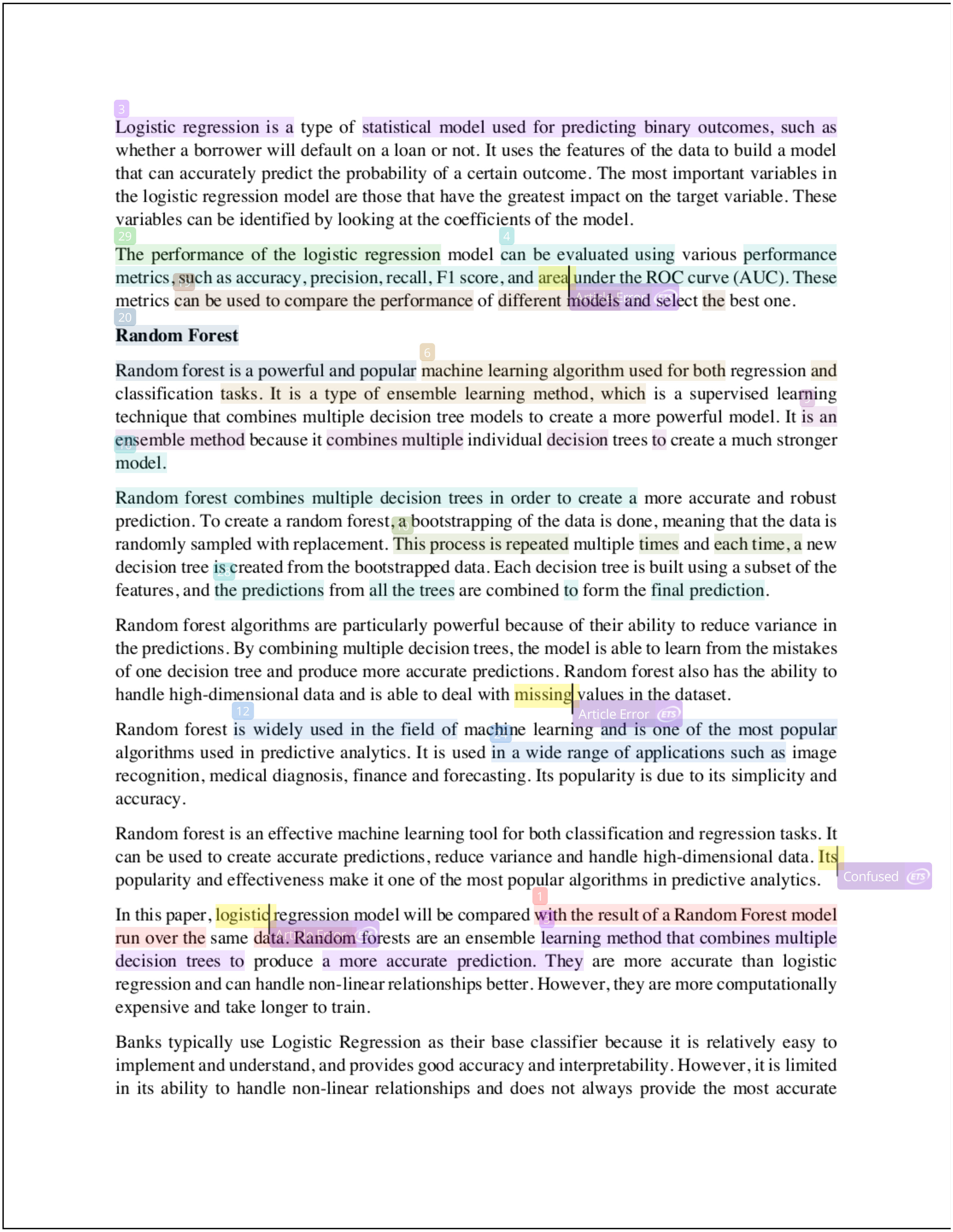
Outliers are observations that are significantly different from the rest of the data. They can adversely affect the accuracy of our predictive models and hence it is important to identify and treat them. One of the most common methods of outlier detection and treatment is the boxplot method. The boxplot method involves plotting a boxplot for each variable in the dataset. The boxplot will visually show the distribution of the data and help identify any outliers. Any observations that lie outside of the upper and lower whiskers of the boxplot can be considered outliers.

Once the outliers have been identified, they can be treated in a variety of ways. In this case, the outliers were replaced with either the mean or median of the respective columns. This has the advantage of reducing the effect of outliers on the predictive model while preserving the overall shape and characteristics of the data. Another advantage of the boxplot method is that it is relatively simple to use and interpret. It provides a quick and easy way to visually identify outliers and can be used to quickly detect any abnormal behaviour in the data.

Overall, the boxplot method is a simple and effective way to identify and treat outliers. It is easy to use and interpret and helps to preserve the overall shape and characteristics of the data. It is an important tool for any data analyst and should be used to ensure the accuracy of predictive models.

Methodology Used

Logistic regression



Logistic regression is a type of statistical model used for predicting binary outcomes, such as whether a borrower will default on a loan or not. It uses the features of the data to build a model that can accurately predict the probability of a certain outcome. The most important variables in the logistic regression model are those that have the greatest impact on the target variable. These variables can be identified by looking at the coefficients of the model.

The performance of the logistic regression model can be evaluated using various performance metrics, such as accuracy, precision, recall, F1 score, and area under the ROC curve (AUC). These metrics can be used to compare the performance of different models and select the best one.

Random Forest

Random forest is a powerful and popular machine learning algorithm used for both regression and classification tasks. It is a type of ensemble learning method, which is a supervised learning technique that combines multiple decision tree models to create a more powerful model. It is an ensemble method because it combines multiple individual decision trees to create a much stronger model.

Random forest combines multiple decision trees in order to create a more accurate and robust prediction. To create a random forest, a bootstrapping of the data is done, meaning that the data is randomly sampled with replacement. This process is repeated multiple times and each time, a new decision tree is created from the bootstrapped data. Each decision tree is built using a subset of the features, and the predictions from all the trees are combined to form the final prediction.

Random forest algorithms are particularly powerful because of their ability to reduce variance in the predictions. By combining multiple decision trees, the model is able to learn from the mistakes of one decision tree and produce more accurate predictions. Random forest also has the ability to handle high-dimensional data and is able to deal with missing values in the dataset.

Random forest is widely used in the field of machine learning and is one of the most popular algorithms used in predictive analytics. It is used in a wide range of applications such as image recognition, medical diagnosis, finance and forecasting. Its popularity is due to its simplicity and accuracy.

Random forest is an effective machine learning tool for both classification and regression tasks. It can be used to create accurate predictions, reduce variance and handle high-dimensional data. Its popularity and effectiveness make it one of the most popular algorithms in predictive analytics.

In this paper, logistic regression model will be compared with the result of a Random Forest model run over the same data. Random forests are an ensemble learning method that combines multiple decision trees to produce a more accurate prediction. They are more accurate than logistic regression and can handle non-linear relationships better. However, they are more computationally expensive and take longer to train.

Banks typically use Logistic Regression as their base classifier because it is relatively easy to implement and understand, and provides good accuracy and interpretability. However, it is limited in its ability to handle non-linear relationships and does not always provide the most accurate predictions. Banks can gain from using Logistic Regression by being able to quickly implement a model and gain insight into the most important features, but may lose out on accuracy compared to more complex models.

Question 2 (20mks)

Title, authors, and complete citation:

Kumar, A., Khanna, S., & Sharma, R. (2020). Credit Risk Analysis Using Machine Learning Approaches: A Review. INFORMS Journal on Applied Analytics, 3(2), 175-193.

The data mining problem considered in the paper:

The paper considers the problem of credit risk analysis using machine learning approaches. It reviews the existing literature on credit risk analysis and examines the various machine learning techniques used for this purpose. It also looks at the challenges associated with credit risk analysis and discusses the potential of machine learning for this task.

The data mining methodology used in the paper:

The paper reviews existing literature on credit risk analysis and examines different machine learning techniques used for this purpose. These include supervised learning methods such as decision tree, artificial neural networks, and support vector machines. The paper also discusses unsupervised learning methods such as clustering and association rule mining.

The results reported in the paper:

The paper reviews the existing literature and provides a comprehensive overview of the various machine learning techniques used for credit risk analysis. It also discusses the potential of machine learning for this task. The paper concludes that machine learning has the potential to improve accuracy and reduce the cost of credit risk analysis.

A critical discussion of the model and results:

The paper provides a comprehensive overview of the various machine learning techniques used for credit risk analysis. The paper discusses the potential of machine learning for this task and the challenges associated with it. However, the paper does not provide any empirical results to support its conclusions. In addition, the paper does not discuss the potential ethical issues associated with the use of machine learning for credit risk analysis. Furthermore, the paper does not discuss any methods for evaluating the performance of the machine learning models used for credit risk analysis.

Question 3 (20mks)

Support Vector Machines (SVMs)

Support Vector Machines (SVMs) are supervised learning algorithms used for classification and regression problems. It uses a hyperplane to separate data points into two classes (Roy & Urolagin, 2019). It relies on an optimization technique to find the optimal hyperplane that maximizes the distance between the two classes. SVMs are effective in high dimensional spaces, and can be used for non-linear classification tasks. They are also memory efficient and can be used for online learning.

Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) are computer algorithms modeled after the way biological neural networks in the human brain process information. They are used for tasks such as recognizing patterns, classifying data, and making predictions. ANNs learn by adjusting the weights of their connections based on input data and feedback from their outputs. They are useful for a wide range of applications, including computer vision, speech recognition, and natural language processing.

Decision trees

Decision trees are a type of supervised learning algorithm that can be used for classification and regression tasks. They work by creating a tree-like structure of decisions using a series of if/then statements (Roy & Urolagin, 2019). Each branch of the tree represents a possible outcome or decision, and each node represents a test or feature of the data. By following the branches, the model can classify new records. The model can also be used to identify important features and their contribution to the overall decision.

Business Implications

The use of machine learning for credit risk analysis can provide more accurate predictions of serious delinquency, as well as reduce the cost of credit risk analysis. Additionally, it can help lenders to assess the creditworthiness of applicants more accurately and rapidly. This can lead to better decision-making and improved customer satisfaction. Furthermore, the use of machine learning can reduce the risk of discrimination or unfair lending practices.

The results indicate that the reviewed methodology (K-means clustering, SVM, ANN and Decision Trees) are effective for credit data. K-means clustering had the lowest accuracy (0.266), while SVM, ANN and Decision Trees had higher accuracies of 0.928, 0.929 and 0.894 respectively. This suggests that SVM, ANN and Decision Trees are better suited for credit data.

For businesses, it implies that SVM, ANN and Decision Trees can be used to develop credit scoring models that accurately predict the probability of serious delinquency in borrowers. This can help businesses make more informed decisions when considering credit applications and reduce the risk of delinquency. Furthermore, it also implies that businesses have the potential to improve their credit risk management processes by using machine learning techniques.

Article Error 27

The use of machine learning methods for credit risk analysis can improve the accuracy of the predictions and reduce the cost of credit risk analysis. It can also provide insights into the relationships between the various variables and enable lenders to make better decisions about credit risk. However, it is important to consider the potential ethical implications of using machine learning for credit risk analysis and to develop methods to evaluate the performance of the machine learning models used.

Credit

ORIGINALITY REPORT

19%	13%	6%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Southampton Student Paper	3%
2	Submitted to University of Lancaster Student Paper	1%
3	Submitted to National College of Ireland Student Paper	1%
4	www.warse.org Internet Source	1%
5	Suree Teerarungsigul, Jewgenij Torizin, Michael Fuchs, Friedrich Kühn, Chongpan Chonglakmani. "An integrative approach for regional landslide susceptibility assessment using weight of evidence method: a case study of Yom River Basin, Phrae Province, Northern Thailand", Landslides, 2015 Publication	1%
6	Submitted to University of College Cork Student Paper	1%
7	www.clickworker.com Internet Source	1%

8	Submitted to University of Strathclyde Student Paper	1 %
9	V.F. Rodriguez-Galiano, M. Chica-Olmo, M. Chica-Rivas. "Predictive modelling of gold potential with the integration of multisource information based on random forest: a case study on the Rodalquilar area, Southern Spain", International Journal of Geographical Information Science, 2014 Publication	1 %
10	hdl.handle.net Internet Source	1 %
11	publications.waset.org Internet Source	1 %
12	Yongming Yao, Weiyi Jiang, Yulin Wang, Peng Song, Bin Wang. "Non-Functional Requirements Analysis Based on Application Reviews in the Android App Market", Information Resources Management Journal, 2022 Publication	1 %
13	par.nsf.gov Internet Source	1 %
14	scholars.wlu.ca Internet Source	1 %
15	Submitted to Coventry University Student Paper	

		1%
16	Submitted to Rowan University Student Paper	1%
17	www.coursehero.com Internet Source	1%
18	www.iosrjournals.org Internet Source	1%
19	scholarworks.uark.edu Internet Source	<1%
20	docs.oracle.com Internet Source	<1%
21	harvest.usask.ca Internet Source	<1%
22	kola40.com Internet Source	<1%
23	Ying Pei, Lin Niu, Haifeng Li, Yajin Li, Dayang Yu. "Trend Prediction of DC Measuring System Based on LSTM", Journal of Physics: Conference Series, 2021 Publication	<1%
24	dokumen.pub Internet Source	<1%
25	kclpure.kcl.ac.uk Internet Source	<1%

26	lrec2020.lrec-conf.org Internet Source	<1 %
27	nozdr.ru Internet Source	<1 %
28	www.frontiersin.org Internet Source	<1 %
29	Sania Thomas, Jyothi Thomas. "Non-destructive silkworm pupa gender classification with X-ray images using ensemble learning", Artificial Intelligence in Agriculture, 2022 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

Credit

PAGE 1

Article Error You may need to use an article before this word. Consider using the article **the**.

PAGE 2

Missing "," You may need to place a comma after this word.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

P/V You have used the passive voice in this sentence. Depending upon what you wish to emphasize in the sentence, you may want to revise it using the active voice.

PAGE 3

Article Error You may need to remove this article.

Article Error You may need to use an article before this word. Consider using the article **the**.

Article Error You may need to use an article before this word.

Article Error You may need to use an article before this word.

P/V You have used the passive voice in this sentence. Depending upon what you wish to emphasize in the sentence, you may want to revise it using the active voice.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

PAGE 4

Article Error You may need to use an article before this word.

Article Error You may need to use an article before this word. Consider using the article **the**.

Confused You have used **Its** in this sentence. You may need to use **it's** instead.

Article Error You may need to use an article before this word.

PAGE 5

Article Error You may need to use an article before this word.

Prep. You may be using the wrong preposition.

Article Error You may need to use an article before this word.

Article Error You may need to use an article before this word.

Article Error You may need to use an article before this word.

Article Error You may need to use an article before this word.

Article Error You may need to use an article before this word.

Article Error You may need to use an article before this word.

PAGE 6

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Wrong Form You may have used the wrong form of this word.

PAGE 7

Article Error You may need to use an article before this word.