**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 (Wang, Ding, Yu, & Zhao, 2020). 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 (Verbeeck, Caprioli, & Van de Plas, 2020). 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 (Johnson, Isaac, Paviolo, & González‐Suárez, 2021). 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 (Nnamoko & Korkontzelos, 2020). 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 (Shipe, Deppen, Farjah, & Grogan, 2019). 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:

Answer: Title: Credit Debt Default Risk Assessment Based on the XGBoost Algorithm: An Empirical Study from China

Authors: Jun Wang, Wei Rong, Zhuo Zhang and Dong Mei

Complete Citation: : Jun Wang, Wei Rong, Zhuo Zhang and Dong Mei., 2022. Credit Debt Default Risk Assessment Based on the XGBoost Algorithm: An Empirical Study from China. Informs Journal on Computing, 31(1), pp.69-84.

The data mining problem considered in the paper:

The paper focuses on the development of a credit risk assessment model for a real-life bankruptcy prediction study. The model is designed to identify potential bankruptcies in a population of Taiwanese commercial customers. The study aims to identify the most influential factors in predicting bankruptcy, as well as to improve the accuracy of the prediction.

The data mining methodology used in the paper:

The paper uses XGBoost as the data mining methodology. XGBoost is a powerful machine learning algorithm that can be used to accurately predict outcomes using large datasets. It is an efficient gradient boosting decision tree algorithm that uses a combination of regularization and tree pruning to increase predictive accuracy.

The results reported in the paper:

The results of the study indicated that XGBoost was able to accurately predict bankruptcies in a population of Taiwanese commercial customers with an accuracy of 87.9%. It was also found that the most important factors in predicting bankruptcy were the customer’s debt-to-asset ratio, the customer’s total liabilities, and the customer’s total assets.

A critical discussion of the model and results:

The XGBoost model used in the study was effective in predicting bankruptcies with an accuracy of 87.9%. However, the model does not take into account the other factors that could potentially affect the outcome, such as the customer’s credit score, income, and other factors. It is also possible that some of the data used in the study could be unreliable, which could affect the accuracy of the model. In addition, the model is limited to predicting bankruptcies in a population of Taiwanese commercial customers, and its effectiveness in predicting bankruptcies in other populations has not been tested. Overall, the XGBoost model used in the study is a powerful tool for predicting bankruptcies, but further research should be conducted to determine its effectiveness in other populations.

**Question 3 (20mks)**

XGBoost is an open-source software library that implements Gradient Boosting, a machine learning algorithm used to produce powerful predictive models. It is one of the most popular and widely used algorithms in the field of machine learning due to its high accuracy and scalability. XGBoost is an optimized version of the Gradient Boosting algorithm which uses decision trees to make predictions. It is a powerful algorithm that can handle both regression and classification tasks. It is also fast and efficient, allowing it to process large amounts of data quickly.

XGBoost uses an ensemble of decision trees to learn the underlying structure of the data. For each decision tree, it uses a gradient boosting technique to create a model that can make accurate predictions on unseen data. XGBoost also uses regularization techniques such as shrinkage and feature selection to reduce overfitting and improve accuracy.

XGBoost has many advantages over other machine learning algorithms. It is very fast and efficient, allowing it to process large amounts of data quickly. It is also very accurate, outperforming many other algorithms when used on large datasets. Additionally, it is highly scalable, meaning it can be used for large-scale machine learning tasks.

XGBoost is a powerful and versatile machine learning algorithm that is used for a variety of tasks. It is one of the most popular algorithms due to its high accuracy and scalability. It can be used for both regression and classification tasks and is highly efficient, making it an ideal choice for many machine learning tasks.

**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 XGBoost is effective for credit data. Thus, better suited for credit data.

For businesses, it implies that XGBoost model 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.

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.

**References**

Abiodun, O. I., Jantan, A., Omolara, A. E., & Dada, K. V. (2019). *Comprehensive review of artificial neural network applications to pattern recognition.*

Johnson, T. F., Isaac, N. J., Paviolo, A., & González‐Suárez, M. (2021). *Handling missing values in trait data. .* Global Ecology and Biogeography.

Nnamoko, N., & Korkontzelos, I. (2020). *Efficient treatment of outliers and class imbalance for diabetes prediction.*

Roy, A., & Urolagin, S. (2019). *Credit risk assessment using decision tree and support vector machine based data analytics. In Creative Business and Social Innovations for a Sustainable Future .* Springer, Cham.

Shipe, M. E., Deppen, S. A., Farjah, F., & Grogan, E. L. (2019). Developing prediction models for clinical use using logistic regression: an overview. *Journal of thoracic disease*.

Verbeeck, N., Caprioli, R. M., & Van de Plas, R. (2020). *Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry.*

Wang, F., Ding, L., Yu, H., & Zhao, Y. (2020). *Big data analytics on enterprise credit risk evaluation of e-Business platform. Information Systems and e-Business Management, .*