Project Deliverable 1 Part 2

ITCS 6190 Cloud Computing for Data Analysis, Fall 2023

Project Group- 5

Team Details:

Member Name	Email ID
Gaurav Avula	gavula@charlotte.edu
Naga Nikhil Bijjala	nbijjala@charlotte.edu
Harshini Karnati	hkarnat1@charlotte.edu
Rohan Katari	rkatari@charlotte.edu
Naga Srivatsav Machiraju	nmachira@charlotte.edu

1. Project Scope and Business Goal: Clearly define the scope and objectives of the project.

o Project Domain:

The primary focus of the project is to address the problem of assessing individuals' creditworthiness with precision using machine learning techniques. The specific tasks and opportunities within this scope include:

a. Credit Score Classification:

The project aims to solve the challenge of accurately determining an individual's creditworthiness. This involves predicting credit scores based on various financial and personal features. Aligns with the AWS Academy Cloud Foundations and Data Engineering goals by incorporating machine learning for predictive analytics, a key aspect of data engineering.

b. Data Collection and Preprocessing:

Gathering relevant data from the provided Kaggle dataset and ensuring its cleanliness and suitability for model training. Reinforces data engineering skills by emphasizing the importance of data quality, consistency, and preprocessing for effective machine learning.

c. Feature Engineering:

Identifying and creating relevant features from the dataset to enhance the model's predictive power. Demonstrates the importance of feature selection and engineering in building robust machine learning models, aligning with the AWS Academy Cloud Foundations and Data Engineering curriculum.

d. Model Selection, Training, and Fine-Tuning:

Choosing an appropriate machine learning algorithm, training the model on the prepared dataset, and optimizing its parameters for better performance. Integrates the machine learning component with the educational goals of AWS Academy Cloud Foundations by emphasizing model selection and optimization.

e. Evaluation and Validation:

Rigorously assessing the model's performance and validating its accuracy and reliability. Aligns with the AWS Academy Cloud Foundations and Data Engineering curriculum by emphasizing the importance of model evaluation and validation for trustworthy machine learning solutions.

f. Integration with AWS Services:

Integrating the developed machine learning model with various AWS tools, such as Amazon SageMaker for deployment, S3 for storage, Lambda for serverless computing, Glue for ETL processes and IAM for access management. Provides practical experience with AWS services, aligning with the AWS Academy Cloud Foundations and Data Engineering program goals.

By addressing these specific problems, tasks, and opportunities, our project aligns with the educational program goals of AWS Academy Cloud Foundations and Data Engineering.

o Domain:

Domain or Industry: Financial Services, specifically Credit Scoring

The financial services industry, particularly within credit scoring, encompasses a complex network of entities and processes that revolve around lending, risk assessment, and financial decision-making. Within this domain, multiple layers and interconnected stakeholders influence and are impacted by credit scoring models.

Key Characteristics:

1. Data-Driven: The financial industry relies heavily on data, and credit scoring is no exception. Lenders assess the creditworthiness of individuals or businesses to make informed lending decisions.

2. Risk Management: Credit scoring is a fundamental aspect of risk management for financial institutions. It helps them evaluate the likelihood of borrowers defaulting on loans, allowing them to mitigate potential losses.

3. Regulatory Compliance: Financial institutions must adhere to strict regulations and compliance standards when handling customer data and making lending decisions. These regulations vary by region and require careful consideration.

Challenges:

1. Data Quality: Ensuring the accuracy and quality of data is essential for reliable credit scoring models. Incomplete or erroneous data can lead to inaccurate assessments.

2. Model Interpretability: Explainability and transparency in credit scoring models are vital to building trust with both stakeholders and regulators. Complex machine learning models can be challenging to interpret.

3. Overfitting: Preventing models from overfitting the data is a challenge. Models must generalize well to make accurate predictions on new, unseen data.

Opportunities:

1. Enhanced Efficiency: Machine learning can streamline and automate the credit scoring process, reducing the time and effort required for assessments.

2. Improved Accuracy: Advanced models can provide more accurate credit risk assessments, reducing the likelihood of bad loans and ultimately benefiting both lenders and borrowers.

3. Financial Inclusion: Machine learning can help expand financial inclusion by providing alternative credit scoring methods for individuals or businesses with limited or no credit history.

Specific Problem/Task: Developing a machine learning model for credit score classification. The primary goal is to accurately assess the creditworthiness of loan applicants, categorizing them into different risk groups, such as high, medium, or low risk.

Stakeholders:

1. Financial Institutions: Banks, credit unions, and lending institutions are the primary stakeholders. They benefit from more accurate credit scoring to make better lending decisions and reduce default risks.

2. Borrowers: Individuals and businesses seeking loans will benefit from a fair and accurate credit assessment, potentially leading to better loan terms.

3. Regulatory Authorities: Regulators oversee and enforce compliance in the financial industry. They benefit from transparent, compliant, and fair credit scoring practices.

4. Data Providers: Credit bureaus and data aggregators play a crucial role in providing the data used for credit scoring. Accurate and up-to-date data is essential for model success.

5. Fintech and Alternative Lenders: Non-traditional financial institutions that use innovative methods to assess credit risk may also be interested in advanced credit scoring models to compete effectively.

Understanding the domain's characteristics, challenges, and opportunities, as well as identifying key stakeholders, is essential for developing a machine learning model for credit score classification that aligns with the needs and objectives of all involved parties.

- ## Literature Review:

1. *L. -l. Zhang, X. -f. Hui and L. Wang, "Application of adaptive support vector machines method in credit scoring," International Conference on Management Science and Engineering, Moscow, Russia, pp. 1410-1415, doi: 10.1109/ICMSE.2009.5317970.*

The assessment of credit scoring has conventionally relied on the intuition and experience of credit managers. However, the emergence of credit classification models, particularly the Support Vector Machine (SVM) classification, has presented a promising avenue for accurately evaluating an individual's creditworthiness. Through a comparative study utilizing Australian and German credit datasets, both SVM and the well-established Backpropagation Neural Network (BNN) methods demonstrated a prediction accuracy of approximately 80%. The study's primary objective was to introduce SVM, a relatively recent learning method based on statistical learning theory, into the realm of credit scoring prediction. By juxtaposing this newer SVM method with the widely adopted BNN, the research highlighted the comparative accuracy of SVM with the established neural network approach. This comparison showcased SVM's potential as a competitive and effective alternative in credit scoring prediction. The findings suggested that SVM stands as a viable tool in this domain, potentially reshaping how credit scoring assessments are approached due to its comparable performance with the traditionally used BNN method.

2. *P. Marikkannu and K. Shanmugapriya, "Classification of customer credit data for intelligent credit scoring system using fuzzy set and MC2 — Domain driven approach," 3rd International Conference on Electronics Computer Technology, Kanyakumari, India, pp. 410-414, doi: 10.1109/ICECTECH.2011.5941782.*

Credit scoring in the banking sector is pivotal for identifying profitable customers while predicting potential bankruptcies. Traditional data-driven methods have often provided imprecise solutions, while domain-driven approaches, like multiple criteria and multiple constraint (MC2) programming, have yielded only satisfying outcomes. To address these limitations, a paper introduces a novel approach using fuzzy sets and domain expertise-driven MC2 programming for classifying customer credit data. This hybrid method employs the knowledge of domain experts to construct linear combinations of attributes, categorizing customers into five classes: best, good, satisfactory, bad, and worst. By using publicly available datasets, the study validates the efficiency of this approach. It focuses on reducing computational complexity through the application of linear combinations of attributes, thereby enhancing system efficiency. The proposed

model aims to identify bankrupt and non-bankrupt customers based on their transaction history, assigning scores by employing the best classifiers associated with each customer. The novelty lies in this domain-driven approach, combining fuzzy sets and MC2 functions to build a model tailored to the banking industry's requirements. The system's major advantage is its ability to derive optimal solutions aligned with the industry's needs, offering a more effective classification of customers into distinct risk categories.

3. *W. Laesanklang, K. Sinapiromsaran and B. Intiyot, "Entropy multi-hyperplane credit scoring model," International Conference on Financial Theory and Engineering, Dubai, United Arab Emirates, pp. 91-94, doi: 10.1109/ICFTE.2010.5499418.*

The Entropy Multi-Hyperplane Credit Scoring Model is a decision-making approach designed to classify loan applicants as payers or defaulters while minimizing misclassification costs. It achieves this by using multiple hyperplanes based on entropy order. In the initial stage, a pair of hyperplanes is created using half of the attributes, which are sorted in ascending order of entropy. These hyperplanes categorize applicants into three groups: payers, defaulters, and unclassified individuals. From the unclassified group, another pair of hyperplanes is generated using the remaining attributes, following the same entropy order. These additional hyperplanes further divide the unclassified group into payers, defaulters, and unclassified applicants in the second stage. Finally, multidimensional hyperplanes created from all attributes are used to classify loan applicants into payers or defaulters. The study develops a mixed-integer programming model for this Entropy Multi-Hyperplane Credit Scoring Model to reduce misclassification costs. Experimental results indicate that this model outperforms a two-stage least-cost credit scoring model in terms of accuracy and requires fewer computational iterations than a multi-hyperplane credit scoring model. Moreover, it exhibits similar performance to other classification methods such as classification trees, neural networks, support vector machines, linear discriminant analysis, and CART. Future research could explore its performance with larger datasets.

4. *Y. Zhuang, Z. Xu and Y. Tang, "A Credit Scoring Model Based on Bayesian Network and Mutual Information," 12th Web Information System and Application Conference (WISA), Jinan, China, pp. 281-286, doi: 10.1109/WISA.2015.31.*

Credit scoring assesses client relationships based on empirical attributes and uses a scoring model to gauge credibility. However, these attributes often harbor uncertainties, necessitating feature selection. Bayesian networks (BN) are adept at handling uncertain

data, and mutual information (MI) serves as a suitable technique for evaluating variable relationships in complex classification tasks. This study introduces BNMI, a credit scoring model that combines BN and MI to address these challenges. The BNMI model employs two core algorithms—BuildBN for variable selection and network learning, and AddParentsToTarget for adjusting the network structure by incorporating relevant attributes. By leveraging the strengths of BN and MI, BNMI aims to enhance credit scoring accuracy. Comparative experiments against three baseline models—decision trees, neural networks, and Bayesian networks—demonstrate that BNMI outperforms these models in terms of receiver operating characteristic (ROC). This suggests a promising application of the BNMI approach in the realm of credit scoring. The model's strength lies in its ability to better handle uncertainties within empirical attributes and improve the accuracy of credit scoring by incorporating the relationship insights offered by BN and MI.

5. *C. R. D. Devi and R. M. Chezian, "A relative evaluation of the performance of ensemble learning in credit scoring," IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, pp. 161-165, doi: 10.1109/ICACA.2016.7887943.*

Credit scoring is a critical focus in banking to detect fraudulent customers and curb illegal activities. Ensemble classifiers in machine learning play a pivotal role in this realm. This study aims to evaluate the accuracy of ensemble methods in classifying customers as good or bad credit risks. Three ensemble methods—AdaBoost, Bagging, and Random Forest—are paired with various learning algorithms and applied to a credit card dataset, preceded by feature selection to identify crucial attributes. The study details specific models: AdaboostDS combines AdaBoost with a decision stump model to strengthen its predictive power. Bagging is employed on J48 and Reduced Error Pruning tree algorithms. These models are tested on the German credit dataset, assessing their performance without feature selection and with different iterations or tree values in the ensembles. The assessment encompasses accuracy, error rate, and ROC curve values. Random Forest consistently demonstrates superior performance in accuracy, ROC values, and minimized root mean squared error compared to the other classifiers. This comparative study provides valuable insights into the effectiveness of ensemble methods in predicting creditworthiness, identifying Random Forest as the most robust performer among the models tested.

o Data Source(s):

We are using the credit score classification dataset available on Kaggle (https://www.kaggle.com/datasets/parisrohan/credit-score-classification). This dataset is segmented into training and testing sets, with the training portion comprising 100,000 rows and 28 columns, while the test dataset consists of 50,000 rows and 27 columns. The primary target column for prediction is the Credit Score, categorized into three main classes: Poor, Good, and Standard.

Within the dataset, information for 12,500 customers is included, encompassing various details such as monthly income, the number of bank accounts and credit cards they possess, interest rates, credit utilization, and more.

The dataset's overall quality is good, offering an extensive array of parameters related to the financial status of each customer. Despite its richness, there are missing values and some columns with blank entries, affecting less than 10% of the customer records. Addressing these issues is feasible through common techniques like replacing missing values with the mean, median, and mode, or deleting rows with substantial missing data that do not significantly contribute to the target variable.

Throughout our exploration of datasets, we encountered challenges in finding one that included detailed financial information. The current dataset primarily covers payment history and monthly account balances, which are crucial factors in determining a person's financial health. In contrast, many other datasets found online lack this level of detail, often providing only basic information like interest rates and the number of accounts.

o Domain Specific Challenges

The domain of our project is working in Financial Credit Scoring and Assessment using Machine Learning. Some identified domain-specific challenges are as follows,

a. Regulatory Compliance:

In the financial domain, there are strict regulations governing the use and handling of sensitive financial information. Compliance with regulations such as GDPR, Fair Credit Reporting Act (FCRA), or other regional financial regulations is crucial.

b. Ethical Considerations:

The use of machine learning in credit scoring raises ethical considerations, particularly regarding fairness and transparency. Ensuring that the model doesn't discriminate against certain demographics is essential.

c. Data Privacy:

Financial data is highly sensitive. Ensuring the privacy and security of individuals' financial information is a critical challenge. Adherence to data protection laws and implementing robust security measures is paramount.

d. Model Explainability:

The interpretability of the chosen machine learning model is essential for stakeholders to understand the factors influencing credit scores. Transparent models may be preferred in this domain to justify decisions and comply with regulatory requirements.

e. Bias and Fairness:

Bias in credit scoring models can have significant social and economic implications. Ensuring fairness in model predictions and addressing biases that may arise from historical data is crucial for ethical credit assessment.

Considering these challenges and addressing them in the development and deployment of our machine learning model will contribute to the responsible and effective use of credit scoring in the financial domain.

o **KPIs ((Key Performance Indicators):**

In the context of building a credit score classification model, several key performance indicators (KPIs) play a crucial role in evaluating the model's effectiveness. Accuracy is a fundamental metric, representing the overall correctness of the model's predictions. However, in the domain of credit score classification, precision and recall are equally vital.

Precision assesses the model's ability to correctly identify positive instances, indicating how well it avoids false positives.

On the other hand, recall gauges the model's capability to capture all actual positive instances, reflecting its sensitivity to true positives.

Striking a balance between precision and recall is often crucial, as an excessively high precision might result in missed opportunities, while a high recall may lead to increased false positives. Additionally, metrics such as the F1 score, which considers both precision and recall and the area under the Receiver Operating Characteristic (ROC) curve are commonly employed in credit scoring models to provide a more comprehensive evaluation.

Ultimately, the choice of KPIs should align with the specific goals and priorities of the credit scoring application, balancing the need for accurate predictions with the consequences of false positives and false negatives in the given financial context.

*****