Credit Card Score Prediction Using Machine Learning

V. Karthik
Student
Department of Computer Science (AIML)
CMR Engineering College,
Hyderabad, India

U. Nagaiah
Assistant Professor
Department of Computer Science (AIML)
CMR Engineering College,
Hyderabad, India

S. Ritesh
Student
Department of Computer Science (AIML)
CMR Engineering College,
Hyderabad, India

T. Esha
Student
Department of Computer Science (AIML)
CMR Engineering College,
Hyderabad, India

V.Arjun
Student
Department of Computer Science (AIML)
CMR Engineering College,
Hyderabad, India

Abstract: In this research a bank loan dataset was utilized to assess the effectiveness of a borrower categorization model and explore machine learning methods. Different models such, as support vector machines, decision trees and random forests were. Their predictive accuracy was compared against regression benchmarks. Performance metrics were analyzed based on a ranking. The findings indicate that the Random Forest model outperforms models. Moreover, the support vector machine model exhibited performance when employing both nonlinear kernels. The study suggests that banks have the potential to enhance their models by incorporating machine learning techniques to create value.

Keywords: *Machine Learning, Artificial Intelligence, Supervised learning, Random Forest, Classification, Regression, TensorFlow.*

I. INTRODUCTION

Credit Card Rating/Scoring:

The concept of "credit rating" refers to evaluating a customer's reliability in meeting responsibilities. The goal is to categorize customers into two groups those deemed trustworthy and those considered risky. It is assumed that responsible customers will meet their commitments while individuals, in the category may struggle to fulfill their obligations.

A credit card is the most common form of payment for a series of goods or services. Features utilized to rate customers based on their credit history. Demonstrates their vulnerability. The rating is comparable to the loan. The point at which decisions are made. Because credibility is primarily the issue of discrimination (good or bad). The rating is traditionally segregated into two primary types based on the tasks and data employed (Bijak and Thomas, 2012). Initially, the the level of detail in the process of applying for a loan. It assesses the candidate's abilities. Requesting the information the employed rate of return for this endeavor is based on the payment frequency of the customer and, later on, the positive or negative situation. Banks must accurately anticipate the potential for customer default. Different periods in order to be profitable (1 month, 3 months, 6 months, etc.). High default risk customers are vulnerable to prohibited from entering the building to allow the bank to take appropriate action to safeguard or regulate itself in order to avoid loss.

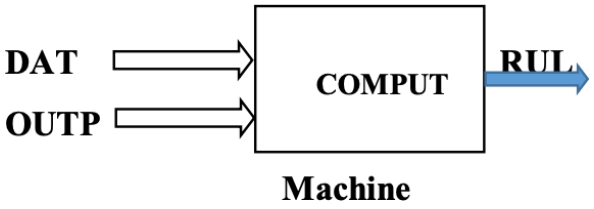
Machine Learning:

The scope of the research is expansive and includes a variety of disciplines including computer science, math, and information. theory, cognitive psychology, and philosophy (AI), McCarthy, et al., 1955. (Cook and Holder, 2001). Intelligent biological processes are often considered to be intrinsic to solutions.

to solutions. The multidisciplinary nature of artificial intelligence technologies, they demonstrate multiple functions strategies and technologies for development (Mira,2008): formal introduction of these concepts, enforcement of the strategies for programming and hardware utilization in these models.

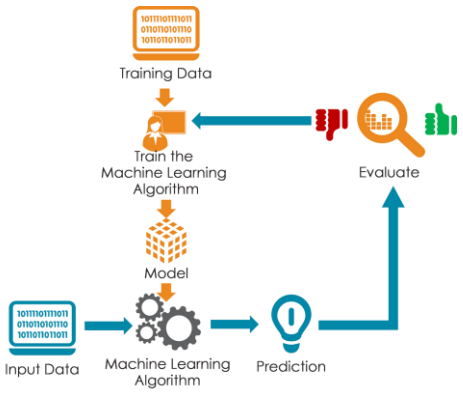
Machine learning primarily revolves around supervised learning, where the goal is to understand the relationship between input and output data. Input data typically consists of descriptions of numerous individual objects, often referred to as examples or instances, while the output is the desired result provided by a supervisor. Classification, a common form of supervised learning, involves segregating different classes through a mapping or discriminative function. These classes represent different outcomes in the data and are termed as class names within the machine learning context. The function responsible for classification is known as a classifier. The training dataset comprises a series of instances with known class names, which are used to define the model's parameters during classification.

These parameters are optimized too accurately map instances to their respective class labels. Machine learning is a revolutionary system that learns from examples, constantly improving itself without direct programming intervention. It operates on the principle that machines can discern patterns from data, such as examples, to generate accurate outcomes, utilizing statistical methods. This innovation has led to practical applications across variousindustries. It shares connections with data mining and Bayesian predictive modeling, employing algorithms to process input dataand generate responses. One notable application is seen in platforms like Netflix, where recommendations are tailored to users based on their viewing history, showcasing thepower of personalized guidance through unsupervised learning. Beyond entertainment, machine learning finds utility in diverse fields, including fraud detection, predictive maintenance, and task automation. Contrary to traditional programming, where rules are explicitly encoded, machine learning algorithms adapt and refine their processes based on data, making them more adaptable to complex scenarios without the need for extensive rule-writing.



How does Machine learning work?

Machine learning serves as the brain where all learning occurs, akin to human cognition. Like humans, machines learn from experience, with increased knowledge enhancing predictive abilities, especially in familiar situations. However, when faced with unfamiliar scenarios, both humans and machines struggle to make accurate predictions. The essence of machine learning lies in learning and reasoning, primarily achieved through pattern recognition and data analysis. As a data scientist, the crucial task is to carefully select and provide relevant data to the machine, forming problem-solving attributes known as feature vectors. These feature vectors serve as a subset of data essential for addressing specific problems. To simplify complex realities and construct models, machines employ sophisticated algorithms. The training phase is pivotal, as it involves organizing and integrating data into the model structure.



As an illustration, consider a scenario where a machine endeavors to grasp the connection between an individual's income and the likelihood of dining at a high-end restaurant. Upon analysis, the machine identifies a positive correlation between a person's salary and their inclination to patronize upscale dining establishments. This outcome forms thebasis of a model.

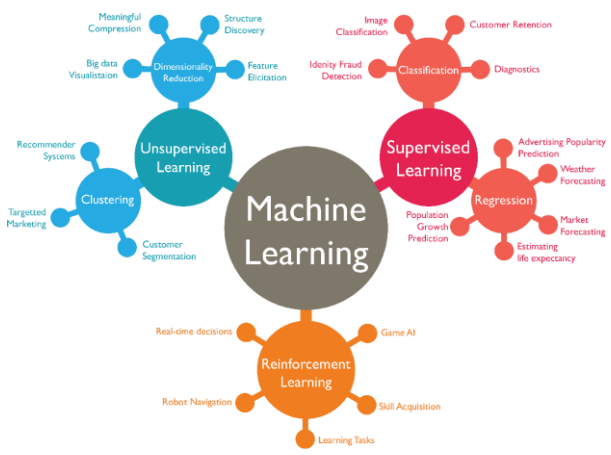
Inferring

Once the model is constructed, its performance can be evaluated using unseen data. This involves transforming the new data into feature vectors, fine-tuning the model, and generating predictions without the necessity of updating rules or retraining the model. Previously trained models can be utilized to derive insights from fresh data.

Summarizing the lifecycle of a machine learning program:

- Problem definition,
- Gathering data,
- Presenting data,
- Training algorithms,
- Testing algorithms,
- Collecting feedback,
- Optimizing algorithms.

Repeat steps 4 through 7 iteratively until achieving a satisfactory outcome. Finally, deploy and utilize the predictive model.



Machine learning encompasses various learning objectives, including supervised, unsupervised, and reinforcement learning. Additionally, there exist numerous algorithms catering to different objectives.

Supervised Learning.

The algorithm identifies the relationship between specific input and desired output through feedback and training data. For instance, professionals can analyze marketing expenses to make deductions. Sales forecasting incorporates weather forecasts. Supervised learning is applicable when the outcome is predetermined, enabling the program to predict new data.

Supervised learning is mainly classified into:

- Classification problem
- Regression problem

Classification:

Let's say you want to predict the gender of a customer for a salesperson. You have access to various data points from your customer database, including height, weight, employment status, salary, shopping history, and more. Each customer's gender is already known to you, either male or female. The objective of the classifier is to evaluate whether a person is more likely to be male or female based on the collected information (features).

Once the model is trained to differentiate between males and females using the available data, it can then make predictions based on new information. For example, if you receive fresh data from an unknown customer and the classifier predicts a 70% likelihood of being male, the computer can infer that the customer is 70% likely to be male and 30% likely to be female.

While there may be multiple classes in a label, in this scenario, there are only two classes: male and female.

However, when the classifier aims to predict an item, it may encounter numerous options (such as glasses, tables, shoes, etc.), with eachoption representing a class.

Regression:
In regression, the focus is on predicting continuous values. For instance, financial analysts might need to predict stock values based on factors like past stock performance and macroeconomic indicators. However, it's important to acknowledge the possibility of errors in these predictions.

Unsupervised Learning:

The algorithm examines input data in an unguided learning process without predefined output variables, such as analyzing customer demographics to detect patterns. This approach can be utilized for data classification or pattern recognition within the algorithm.

Algorithm Name	Description	Type
Linear regression	Allows for each characteristic to be linked to the outcome for forecasting upcoming values.	Regression
Logistic regression	Linear regression extension adapted for classification purposes, where the output variable is binary rather than continuous, such as distinguishing between two categories like black and white, as opposed to dealing with an infinite range of possibilities like colors.	Classification
Decision Tree	A model, either for classification or regression, that divides data-feature values into branches at decision points (for instance, if a feature represents color, each possible color creates a new branch) until reaching a final decision output.	Classification Regression
Naive Bayes	A model for classification or regression that partitions data-feature values into branches at decision nodes, where each potential value of a feature, such as color, creates a separate branch, ultimately leading to a final decision output.	Classification Regression
Support Vector Machine	Support Vector Machine (SVM) is a widely used method for classification tasks. SVM identifies a hyperplane that effectively separates classes and performs optimally when paired with a nonlinear approach.	Regression(not common) Classification

Algorithm Name	Description	Type
Random forest	This method relies on a decision tree as its foundation, significantly enhancing accuracy. In a random forest, multiple simple decision trees are constructed, and a label is selected for prediction based on the "majority vote" principle. The ultimate classification predictionis determined by the label with the highest number of votes, while for regression tasks, the final prediction is calculated as the average forecast generated by all the trees.	Classification Regression
AdaBoost	An approach for classification or regression that evaluates different models based on their predictive accuracy before selecting the most suitable one.	Classification Regression
Gradient-boosting trees	Gradient boosting trees represent an advanced technique for both classification and regression tasks. This method prioritizes addressing the errors made by previous trees to refine its predictions.	Classification Regression

II. APPLICATION OF MACHINE LEARNING

Augmentation:

- Machine learning offers assistance in resolving everyday personal or business challenges without necessitating complete control over the outcomes. It finds applications across various domains such as virtual assistants, data analysis, and software solutions, thereby mitigating human errors.

Government:

- Governments employ machine learning for overseeing public safety and administering government services. For instance, in countries like China, there's significant utilization of facial recognition technology. Artificial intelligence is leveraged to proactively identify and prevent potential disturbances or disruptive individuals from emerging.

Finance Industry:

- The financial sector increasingly adopts machine learning technologies. Primarily, banks utilize these technologies to identify patterns within data and to mitigate instances of fraud.

Automation:

- In various domains, machine learning operates autonomously, independent of human intervention. For example, robots are deployed in manufacturing plants to execute crucial tasks without human involvement.

Healthcare industry:

- Healthcare stands out as one of the pioneering industries to integrate machine learning alongside image recognition technology.

Examples of the application of machine learning in the supply chain:

Machine learning delivers impressive results in visual pattern recognition, opening up numerous new applications for physical inspection and maintenance throughout the supply chain. Unsupervised learning enables rapid identification of similar patterns across different datasets. During transportation, machines can conduct comprehensive quality inspections on the condition and wear of entire logistics centers. For instance, IBM Watson platform can detect damage on shipping containers by integrating vision and system data for real-time tracking, reporting, and recommendations. Recently, reservemanagers have increasingly relied on advanced prediction methods combining big data and machine learning, resulting in 20% to 30% higher accuracy compared to traditional tools, potentially reducing storage costs by 2-3%.

An exemplary illustration of machine learning in action is Google's self-driving car. Equipped with a roof full of lasers for positioning relative to the environment and a front radar for detecting surrounding vehicle speed and movement, the car can not only navigate autonomously but also anticipate potential driver behaviors. Processing nearly 1 GB of data per second, these technologies enable efficient and safe autonomous driving.

Deep learning:

Deep learning is a type of software that emulates the neural network of the human brain. It falls under the umbrella of machine learning, specifically focusing on deep neural networks. The data undergoes multiple layers of examination within the machine, with the depth of the model determined by the number of layers. Deep learning is a relatively new concept within the field of artificial intelligence, and its learning process is facilitated through neural networks.

Reinforcement Learning:

Reinforcement learning, a branch of machine learning, involves training systems through the receipt of virtual " rewards" or "punishments," essentially through trial and error. Google Deep Mind employed reinforcement learning to surpass the human champion in the game of Go, and it enhances gaming experiences by creating more intelligent robots in video games.

One of the most famous algorithms:

- Q-learning
- Deep Q network
- State-Action-Reward-State-Action(SARSA)
- Deep Deterministic Policy Gradient(DDPG)

Applications/ Examples of deep learning applications:

AI in Finance:

The financial technology (fintech) sector has embraced artificial intelligence (AI) to enhance efficiency, lower expenses, and enhance value. Deep learning revolutionizes lending processes by enhancing credit worthiness assessment and utilizing AI to more accurately evaluate risks and the qualifications of job applicants. Underwrite, a fintech company, offers AI solutions to lenders, which are fundamentally more effective than conventional approaches.

AI in HR:

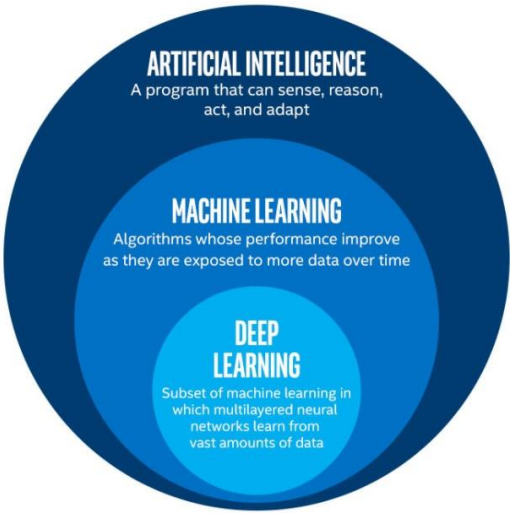
Under Armour, a sportswear company, is revamping its recruitment process with the help of artificial intelligence (AI) innovations. Specifically, Under Armour managed to slash its retail store order time by 35%. The company's popularity surged in 2012, resulting in an average influx of 30,000 resumes per month. However, sifting through these applications and initiating the selection and interview procedures proved to be time-consuming. This prolonged recruitment process hindered Under Armour's ability to adequately staff its retail outlets. Although the company possessed various HR technologies.

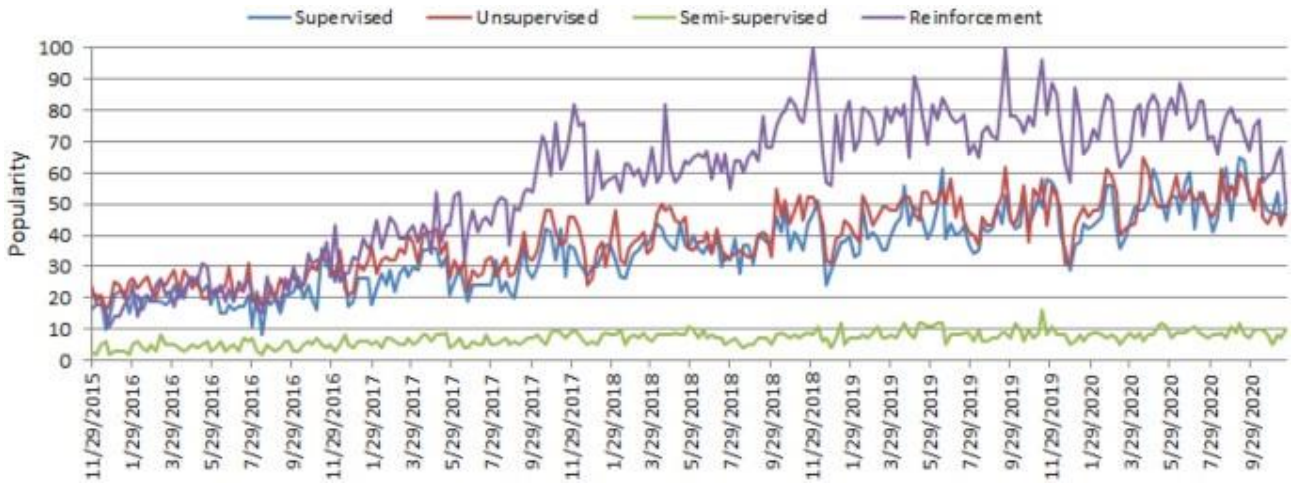
Difference between Machine Learning and Deep Learning:

	Machine Learning	Deep Learning
Data Dependencies	Outstanding outcomes were attained on a modest to moderate dataset.	Excellent performance with large amount of data.
Hardware dependencies	Work on a low-cost computer.	DL requires a powerful processor, preferably one with a GPU because it conducts alot of matrix multiplication.
Feature engineering	The characteristics that represent the data must be understood.	There's no need to figure out which feature best describes the data.
Execution time	From few minutes to hours.	Several weeks are possible. A large number of weights must be computed using a neural network.
Interpretability	Some algorithms are simple to understand (logistic, decision tree), while others are nearly hard to comprehend (SVM, XGBoost)	Difficult to impossible

In contrast to deep learning, machine learning necessitates a smaller amount of data for algorithm training. Deep learning, on the other hand, mandates a substantial and varied dataset to establish its framework. Furthermore, machine learning boasts quicker model training, typically spanning from a few days to a week. Deep learning holds an edge over machine learning in terms of accuracy, as it doesn't require prior knowledge of the most relevant data features. Neural networks have developed the ability to identify crucial characteristics independently. In machine learning, however, selecting the features to incorporate into the model is essential.

Artificial Intelligence:





TensorFlow:

TensorFlow, developed by Google, stands out as one of the most renowned deep learning libraries globally. It's extensively employed across various Google products to enhance search engines, translations, headlines, and recommendations through machine learning. For instance, Google implements AI-driven search enhancements, providing users with word suggestions as they type in the search bar, aiming to improve the speed and efficiency of searches for Google users. Google aspires to leverage its vast datasets to empower both machines and users to enhance their overall experience.

Different groups use machine learning:

- Researchers
- Data scientists
- Programmers

Collaboration using shared tools can significantly enhance efficiency for all involved. Google, boasting the world's largest computer infrastructure, plays a pivotal role in the development of TensorFlow. This deep learning library, primarily created by the Google Brain team, is tailored to accelerate machine training and the exploration of deep neural networks. TensorFlow offers versatility by being compatible with various CPUs or GPUs, including those found in mobile systems. It supports multiple programming languages, such as Python, C++, or Java, through various containers.

TensorFlow Architecture:

Tensorflow architecture works in three parts:

- Preprocessing the data
- Build the model
- Train and estimate the model

TensorFlow derives its name from its primary input format, which consists of multi dimensional arrays known as tensors. These tensors are processed through a flowchart, or graph, starting from an entry point, passing through various operations, and exiting at the other end. This versatile system earned its name because tensors enter, undergo operations and exit on the opposite side.

TensorFlow has the capability to be trained on multiple machines and executed on a different machine once the model is trained. Both GPUs and CPUs can be utilized for training and running the model. GPUs, initially developed for gaming purposes, were found to excel in matrix operations and algebra, making them highly efficient for tasks like deep learning, which heavily relies on matrix multiplication. TensorFlow, written in C++, is particularly adept at computing matrix multiplication, contributing to its speed and efficiency. Despite being implemented in C++, TensorFlow can be interacted with and controlled using various programming languages, with Python being notably popular.

III. MAIN CONCEPTS

This segment presents a formal description of the components associated with supervised learning. In a standard supervised learning setup, a training set *S* consisting of examples $x \in X$ and corresponding output values $y \in Y$ is provided. Here, *X* represents the set containing all possible examples in the input space, denoted as $X=\{x1,x2,x3....\}$. Typically, each example *x* is described using a vector of features or attribute values. In the context of machine learning terminology, a function can be considered as one of two types of data.

- number: the feature value is a real number;
- categorical: the feature value is a member of a predetermined finite set.

The statistical text is different in the expansion of data types, including:

- Nominal value: the characteristic value is a member of an unordered set, such as {tenant, owner, other}.
- ordinal: the characteristic value is ordered The members of the set S. {high, medium, low}.
- Interval: the characteristic value is measured in a fixed and equal unit, and is a member of an ordered set, for example The temperature is in degrees Fahrenheit.
- Ratio: The characteristic value has the attribute of the interval data type, but has an absolute zero point (that is, no negative value). For instance, Income to expenditure. It is a collection of all possible outcomes. The value training setS in the output space consists of n tuples (or instances).

$$S = \{(x_1, y_1),. , (x_i,y_i),. , (x_n, y_n)\}$$

It's crucial to emphasize that the examples *x* within the training set *S* are presumed to be independent and originate from the same distribution as the elements of *X*. Here, *X* represents an unknown but consistent probability distribution function denoted as $P(x,y)$.

The effective operation and enhancement of the rating system are key concerns for banks, customers, and regulatory bodies, especially considering the latest developments in the global economy.

MINIMIZING RISK:

To determine the optimal classifier within the hypothesis space, the loss function serves as a quantitative evaluation of the relationship between the predicted outcome $h(x)$ and the expected result *y*. The ideal function *h* minimizes the expected error or risk.

$$R(h) = L(h(x), \quad P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)}$$

Where L represents the correspondingly selected loss function. For binary classification, the loss function is usually 0/1 loss. If $y = h(x)$, then $L(h(x), y)$ is 0, otherwise it is 1. (X, y) is unknown, and the risk cannot be minimized directly, but the solution is close to the expected minimumerror of the available training set S. and is discriminatory. Classification basis (Cunningham et al.) The generation-based method checks common $P(x,y)$ *or* $P(y|x)P(x)$ probability models, and uses Bayes' theorem to obtain the required posterior probability.

To determine the optimal classifier within the hypothesis space, the loss function serves as a quantitative evaluation of the relationship between the predicted outcome $h(x)$ and the expected result *y*. The ideal function *h* minimizes the expected error or risk.

IV. SUMMARY

Banks utilize credit ratings to categorize and evaluate the risk associated with potential loss stemming from changes in customers' financial capabilities and willingness to meet their financial obligations. Initially, the assessment of private clients' credibility was subjective in nature.

A scorecard is a digital tool employed to assign levels to customer attributes generatinga risk value indicative of the likelihood that a customer may default on their financial obligations.

The process of developing a scorecard comprises three primary stages: (i) compiling a dataset, (ii) modeling, and (iii) documentation.Various methods are employed during dataset creation and modeling. Developing a scorecard is a meticulous process that necessitates consideration of multiple factors. Demographic shifts and economic fluctuations can introduce diverse scenarios, rendering traditional standards and dashboard design approaches inadequate.

One common challenge arises when thedataset contains limited default values, posing difficulty in constructing a reliable scorecard. Moreover, privacy laws and corporate sensitivities often impede access to actual credit information, presenting a challenge for researchers. Artificial data can help overcome these limitations, enabling scientists to create specific conditions for studying particular issues.

REFERENCES

1. Jayagopal, B. “Applying Data Mining Techniques to Credit”
2. Hand, David J., and William E. Henley. “Statistical classification methods in consumer credit scoring: a review." Journal of the Royal Statistical Society: Series A (Statistics in Society) 160.3 (1997): 523-541.
3. Wallis, Mark, et al. “Credit Rating Forecasting Using Machine Learning Techniques.” Advances in Data Mining and Database Management Managerial Perspectives on Intelligent Big Data Analytics, 2019, pp. 180–198.
4. Weston, Liz Pulliam. Your Credit Score: How to Improve the 3-Digit Number That Shapes Your *Financial Future*. FT Press, 2016.
5. Understanding Your Credit Report and Credit Score. Financial Consumer Agency of Canada, 2012.
6. Li, Xiao-Lin, and Yu Zhong. "An overview of personal credit scoring: techniques and future work.”(2012).
7. West, Jarrod, and Maumita Bhattacharya. "SomeExperimental Issues in Financial Fraud Mining.” Procedia Computer Science 80 (2016): 1734-1744.
8. Li, Xiao-Lin, and Yu Zhong. "An overview of personal credit scoring: techniques and future work.” (2012).