|  |
| --- |
| /Users/apple/Library/Containers/com.microsoft.Outlook/Data/Library/Caches/Signatures/signature_2105988401Close-up image showing the leaf-sides of two oversized books side-by-side on a bookshelf, with additional books in soft focus background |
| Artificial Intelligence  Submitted to: Dr. Rami Ibrahim |
| |  |  |  | | --- | --- | --- | | Submitted by: YOUSEF ABU ALI | Student number: 21110294 | Course number :10204250 Year: Spring 2022-2023 | |

The foundations of AI and the development of an AI system

1. Part 2:
   1. Introduction:

The problem I am trying to solve is the credit score classification depending on people’s past credit related information. The goal of my project is to build a machine learning model that could accurately classify the clients credit score brackets in a bank into (poor or Good). This project aims to remove the human effort and provide technologies that could effectively discover worthiness of client’s card, and to know what decisions to make, such as knowing wither to approve lending money to a client person or not.

Some of this project benefits are:

* Improving the efficiency of the organization’s work by applying a technology in their system that could reduce the amount of time and effort wasted on human’s evaluation for client’s credit score, which could help them concentrate on other important aspects.
* Improving the accuracy of the decisions taken by applying machine learning, the project can reach higher accuracy considering the credit score classification especially when compared to the traditional methods.

Considering the novelty contribution of my project, it mainly focuses on applying machine learning algorithms in order to provide automatic scoring for the client’s credit, as well as to improve the overall process of credit score classification. Using my project, I could have enhanced the prediction accuracy of the credit score by applying different methods of machine learning. Moreover, my project contributes to making the process of credit scoring fully automated, as well as reducing human efforts and errors which mainly increases the efficiency of the organization’s work.

* 1. Deployment
     1. Materials:
* **What is the Source/description of the dataset?**

I found this dataset on Kaggle website on the following link: (<https://www.kaggle.com/datasets/parisrohan/credit-score-classification?select=train.csv>) . It was written that Rohan Paris is the (Owner) of the dataset. The general description that was written on the dataset was written in a scenario ( you are a data scientist working in well known finance company that has collected some bank details on the client’s credit information, and you are tasked to build machine learning models that is able to find the credit score of a client by having the credit related information of that client).

* **How was the dataset collected:**

This dataset was collected by one of the workers in an unknown global finance company (Bank). It was collected from clients past information, in order to try applying some machine learning models on it (as described) in order to find out wither it is possible to relay on the technologies to start making decisions on the credit score of a client, instead of wasting employees time on it.

* **The dataset attribute:**

1. ID: which represents the special number for a row of a client information.
2. Customer\_Id: represents the unique ID for the clients in the bank, as a unique Identification number.
3. Month: it shows the month of the year in which that information was taken.
4. Name: represent the name of the client.
5. Age: it shows the client’s age.
6. SSN: it shows the social security number of a client.
7. Occupation: it shows the clients occupation`.
8. Annual\_Income: it shows the income of a client.
9. Monthly\_Inhand\_Salary: it shows the monthly base salary of a client.
10. Num\_Bank\_Accounts: shows the number of bank accounts for a client.
11. Num\_Credit\_Card: it shows the number of credit card that a client has.
12. Interest\_Rate: it shows the client’s interest rate of their credit card.
13. Num\_of\_Loan: it shows the n umber of loans that a client has taken from the bank.
14. Type\_of\_Loan: it shows the type of loans that a client has taken.
15. Delay\_from\_due\_date: it shows the average number of days that client has delayed from paying.
16. Num\_of\_Delayed\_Payment: it represents the number of payments that has been delayed by a person.
17. Changed\_Credit\_Limit: it show a percentage of the change of the limits of credit card.
18. Num\_Credit\_Inquiries: it shows the number of credit card inquiries.
19. Credit\_Mix: it shows a classification of a mix of credits.
20. Outstanding\_Debt: it shows the remaining dept that needs to be paid by a client in USD.
21. Credit\_Utilization\_Ratio: it shows utilization ratio of the clients credit card.
22. Credit\_History\_Age: it shows the age of a clients credit card history.
23. Payment\_of\_Min\_Amount: it shows wither only the minimum amount was paid by a client.
24. Total\_EMI\_per\_month: it shows the monthly payments of EMI in USD.
25. Amount\_invested\_monthly: is shows the amount of money invested by a customer monthly in USD.
26. Payment\_Behaviour: It shows the payment behavior of a client in USD.
27. Monthly\_Balance: it shows the monthly balance amount of a client in USD.
28. Credit\_Score: it is the feature we are trying to find, it shows the over all credit score of a client ( Poor, standard, Good ).

* **What is the size of the dataset?**

This dataset contains 57864 rows and 28 columns.

* **In how many projects was the dataset used?**

As written on the website, this dataset was used by 61 people, probably data scientists who tried to find the best optimal solution for this problem.

* **How was the dataset pre-processed?**
* In the beginning I started by reducing the data set because I only wanted to find the result of either a credit score is classified under poor or good, therefore I dropped the rows that contains a credit score of standards.
* After that I applied the Label encoder in order to encode the categorical variables.
* I filled the null values in the features that has null values with either the mean or the mode.
* I dropped the features that wasn’t considered important.
* I dropped the features that had low correlation with the dependent variable.
* I used the get dummies to encode the categorical variable Occupation.
* **The exploratory analysis I have made to understand the data:**
* First, I used the Score.info() method in order to get general information on the data, such as number of null values, and the data type for the features.
* I used the bar plot for the dependent variable Credit\_Score in order to get a general Idea of the frequency for the poor and good score. It showed that poor scores were more frequent.
* I used the bar plot for the Occupation in order to get a general Idea of what clients most often takes loans from the bank. It showed that scientists were the most clients that take loans.
* I used the bar plot for the month features in order to get a general Idea of what was the most active month in loans taking. It showed that April and June were the most active months.
* I used the Correlation Matrix Heatmap in order to get an idea of the correlation between all the variables with the dependent variable. And based on it I removed features with low correlation.
* I used the describe() method in order to get information on the features such as the mean, minimum and maximum value for each feature.
  + 1. Methods:
* **The models I selected for my project:**

In my project I used the Artificial neural networks as well as the decision tree as machine learning algorithms.

* Artificial neural networks:

Artificial neural networks are also known as neural networks. It is a part of machine learning, which provides the foundation of deep learning algorithms. The name neural network was influenced by the human brain. Therefore, they simulate the way neurons in the human brain communicate with each other. (IBM, 2021)

Artificial neural networks are made up of layers called (nodes). They include the input layer, hidden layers, as well as output layers. Each one of these nodes has a specific wight and threshold, and each one of them is connected to the other layers. If any of these layer’s output exceeds the threshold it has, it will be triggered and it will send the data to the next node (layer). Neural networks are able to learn, and they can improve their accuracy through learning on the data.

* Decision tree classifier:

The decision tree is considered as a supervised learning algorithm (uses labels). This algorithm good be used for both classification as well as the regression models. This algorithm looks like a tree In the way it learns, it contains root nodes, internal nodes (decision nodes), as well as leaf nodes. ((Coursera, n.d.))

* **Why did I select these models:**
* ANN

I used the ANN for the amazing process and benefits it offers. ANN models are excellent when it comes to recognizing patterns, even if a complex data set was used, and it is able to models which are considered non-linear.

Moreover, ANN model is able to learn, change, and improve from the training data, which helps them in making more accurate predictions on new data.

ANN is also able to perform multiple processes at the same time, which gives them the advantage of being able to handle huge amounts of data effectively.

These ANN advantages are useful especially in tasks such as pattern recognition, data analysis, as well as making predictions in artificial intelligence and data science fields. (Simplilearn.com, n.d.)

* Decision tree:

I used the decision tree algorithm because it provides an excellent method when making judgments since they state the problem and all the possible solutions .

Decision trees is an important machine learning algorithm as it’s known for its flexibility and effectiveness. It is able to work with multiple data types such as numerical data, and categorical data, without needing to perform a lot of preprocessing.

Moreover, decision trees are able to with outliers as well as working with huge data sets.

* **Drawing pipeline (architecture) for each model:**
* Model 1:

A picture containing text, screenshot, font, line

Description automatically generated

* Model 2:

A picture containing text, screenshot, font, line

Description automatically generated

* **Explaining the technical implementation for each model:**

I will start first by explaining the common technical implementation:

At the bigging I started by loading the dataset into the python notebook using the pandas data frame. I printed some information about the data set such as the size of the data set, using .shape, the number of null values in the data set using isnull().sum() functions, and the data type for each feature using the .info() function.

After that I dropped the score 2 from the dependent variable, to only keep 1 and 3, which are the (Poor or Good) results. And then I performed the Label Encoder on the dependent variable in order to turn 1 and 3 to 0 and 1. Where 0 represents a Poor score, and 1 represents a good score. I also applied the Label Encoder to the (type of loan) feature.

After that I started plotting some visualizations for some features to gain some information on them. I started by the dependent variable (credit score) and I applied the bar plot on it to get a general Idea of what was the most common credit score. I realized that the score 0 (Poor) was more common than (Good) score. But it wasn’t that much of a difference therefore I didn’t apply the over sampling or under sampling methods.

I also applied the bar plot on the occupation and month features. It showed that the most common occupation was scientists. And the most common months were April and June.

Next for the missing values, I filled the float data types features with the Mean. And I filled the object data type features with the mode.

After that I used the correlation heat map in order to know the correlation between the dependent variable with the independent variables. After that I dropped the unimportant features and the features that has a low correlation with the dependent variable.

And before I started applying my models, I split the data into training and testing. And then I used the select K best features selection method, and I chose the K value as 19 features.

* Model 1:

My implementation for the ANN model included several layers connected in a sequence way.

I used the Keras library which helps in the building and training of the neural networks. In the bigging I used the Sequetial() function in order to initialize the model. Which helps in creating an empty model that we will use to add layers to that model.

After that I added the input layer to the model, which contains 6 neurons and uses an activation function ReLU, this helps include non-linearity to the model network. After that, I added another layer with 6 neurons .

Which was followed by an output layer with 1 neuron. This layer has the sigmoid activation function which helps in binary classification. Next, I specified the optimizer as the Adam optimizer, and the loss function as the (binary\_crossentropy), and the accuracy for the evaluation metric.

Next, I trained the model by using the training data and the target variables. I set the batch size as 10, which is the number of samples. And the epochs as 20, which is the number of times the training data set will be passed into the model.

After that, I used the predict() function, in order to make prediction on the test data, which make the model make probabilities for each sample.

At the end I used the threshold number as 0.5 to the probability’s prediction. Any value over 0.5 is positive and below or equal will be negative.

* Model 2:

In the begging I started by creating an object for the decision tree classifier using the (DecisionTreeClassifier()).

In order to access the performance of the Decision Tree Classifier I initialized variables such as the accuracy, precision, recall, and f1-score. I will take the average for these metrics after multiple iterations to provide an evaluation for the overall iterations.

Each one of the loops in the for loop will perform the following steps:

1. Splitting the data:

In each iteration the dataset will be randomly splitted into training and testing suing the train\_test\_split () function.

1. Training the model:

In each iteration the classifier of decision tree will be trained on the training data by using the function fit(), which helps in analyzing the features and the labels of traing portion, in order to build the decision tree model.

1. Evaluation metrics:

In each iteration the evaluation metrics will be calculated for the same loop and stored to be used later.

* **Explaining how the models could work together with other models/tools/approaches in the organization?**

My project can have a lot of benefits regarding the organization, which is the bank:

* This project could be connected to the bank’s system in order to perform automatic preprocessing and gathering for the data. this is considered as a guarantee that the models could have real time access to the newest and most recent data, which is essential to make better and more accurate predictions.
* Improving data quality would lead the models to have better prediction ability, by adding to credit score data set additional financial information, such as the transactions data as well as the clients’ personal information. Adding these data could help models capture better patterns and have better correlations.
* We can somehow merge the two predictions models with each other such as the ANN and the Decision tree. This method might outperform the individual models by connecting the strengths from each method. If we think about it, the ANN is excellent in working with non-linear relationships, and the Decision tree is able to deal with categorical data in a more effective way. Therefore, if we connected them together, we might have better results. We might merge the two models to make the system choose the model that gives better accuracy in a certain task. Also, we can divide the data to numerical and categorical and make each model work on one of them, and then combine the results. (Brownlee, 2021)
* Applying these models for predicting credit score for clients would be very beneficial for the owners. First of all, it would help them gain more profits by only having to deal with loyal and trustworthiness clients. As well as, cutting the numbers of workers in the bank by making machines responsible for such tasks.
* Even if the bank won’t rely totally on this technology, it could work greatly with employees to help them save time, effort, and make better decisions.
* **The measures of performance I used to evaluate my model and rationale for using such metrics.**

For both models I used accuracy, recall, f1-score, and precision in order to evaluate my model’s performance.

Using these evaluation metrics is beneficial due to the following:

* Accuracy: the accuracy give a general idea of how the model is performing ( in making correct predictions) over all. In other words, it tills how much this model is good in prediction the problem you are having.
* Recall: the recall find how much the model can find the cases that are positive correctly. In my case it means identifying how accurately the model is identifying clients with a good credit score. (Google, 2019)
* Presision: focuses in finding how the model is perfoming in making positive predictions. (Google, 2019)
* F1-score: it solves the problem between the recall and precision, because if one of them increases the other will decrease. Therefore it is good to make a balance between them. and it is considered as a affective measure to find the over all models performance.
* **Evaluation on how based on the performance measures I was able to enhance the model?**

The first time I applied the models on my dataset, I had an accuracy of 0.20 for the ANN model. And an accuracy of 0.45 for the decision tree model. Therefore, I applied certain techniques to enhance my models. The mistake I did at the bigging is that I removed certain features because I felt based on their description that they were not important. Also, I tried avoiding the feature that contained categorical variables, and I tried the numerical variables only. Which made the accuracy even worse. But then I found the solution:

* First of all, I started by encoding the categorical variables with different encoding methods.
* Then, I filled the null values with the appropriate methods, such as filling numerical variables with the mean, and the categorical variables with the mode.
* After that I applied the correlation heat map on my dataset. I remove all the features that had a low correlation with the dependent variable.
* I removed the features that were unrelated to the prediction itself, such as the ID, SSN, client name, and the customer Id, as they are related only to the customer.
* I applied a feature selection technique to help me choose the best features for my model. I used the select KBest method. And I tired multiple values for K until I choose the best one, I found which was 19 features.
* For the ANN model I tried to change the batch\_size multiple of times to see how the accuracy gets affected.
* For the decision tree I used a for loop to find the best possible result, and I tried changing the number of iterations for it, and I stayed at 10 iterations.

After each one of these techniques, I was monitoring the accuracy of my models and wither they were getting better or worse. But I found that each technique I used had increased the model’s accuracy in a clear way.

* + 1. Results and Discussion:
* **Analysis of the results.**

For the ANN model:

My ANN model was trained 20 times. It started with an accuracy of 0.78 and loss of 0.4510. and through the iterations the model’s accuracy started going up and the loss function was decreasing gradually. The final accuracy of the model was 83.37 .

For the decision tree:

The model was trained and evaluated for 10 iterations. En each iteration the data was split in a different way. But over all the mean ( average) for the model’s accuracy was 90.80.

For the first model (ANN) the Recall was (0.84), Precision (0.82), F1-Score (0.82), and an Accuracy of (0.83).

For the second method (Decision tree) the Recall was (0.87), Precision (0.88), F1-Score (0.87), and an Accuracy of (0.90).

Overall, the Decision tree has giving a better performance than the ANN model in all performance metrics. The decision tree has given better results for the recall, precision, f1-score, and accuracy. Therefore, the decision tree shows better capabilities of dealing with the credit score data set, and a general better performance when it comes to classifying the credit scores for the clients. However, this doesn’t mean that the ANN can’t be used, as the performance results were not that far than the decision tree.

* **The Possibility of implementing this project in Jordan.**

Based on the amazing results for the models, I think that this project would be very important in Jordan. As the banks in Jordan could use these models to help them make better prediction when it comes to evaluating the credit score for the clients.

Moreover, I think that implementing such project will be beneficial for saving employees times as well as focusing on more important things.

In general, I think that applying this project for credit scoring in Jordan banks is not only limited to its accuracy and efficiency. I would also be beneficial for the development and growth of the banking sector.

By applying the amazing power of machine learning, banks will be able to make more stable and enhanced decisions, as well as reducing any risks, and they can present some valuable solutions for their clients.

* **What further enhancements can be made in the future?**

The first thing that comes to mind is to try all the possible machine learning models in order to enhance the credit scoring process. By employing different machine learning models, I can assess the performance for each one using certain performance measures, and then I can choose the best models for my project, which will make the credit scoring method better.

Moreover, I can try multiple feature selection methods in my project, and I can apply each feature selection method on all the machine learning models that I will try. Which later on, I choose the best results that comes from the best machine learning model with the best feature selection technique.

One of the limitations that might have stopped me from getting better results in my project is probably that I hadn’t had enough (deep) information about the features in the dataset. Moreover, I might have some limitations when it comes to applying difficult techniques such as feature engineering.

If I had better knowledge about the features in the data set, and good knowledge in feature engineering I might be able to transform some features and create new features from existing ones.

* **What is my role in building/improving this project?**
* As I discussed before in my report, I would like to extend the information on the data set. Such as increasing the size of the data set. Or even adding more features to the data set which might help me in getting better results in my machine learning algorithms.
* Also, I might change the training for the data, such as changing the ratio for training and testing to see wither the result get better or not.
* I also might try to change some specifics on the ANN model, such as adding more layers, or changing the batch size or epochs to compare the performance wither it gets enhanced.
* I also might try changing the number of K in the KBest feature selection method.
* I might try adding some extra preprocessing techniques, such as applying the min-max normalization.
* I might search for outliers in the data set and try removing them if they exist.
* I might try dropping rows that contain null values instead of filling them, as their number is very small considering the size of the data set.
  1. References:

1. IBM (2021). What are Neural Networks? | IBM. [online] www.ibm.com. Available at: <https://www.ibm.com/topics/neural-networks>.
2. Coursera. (n.d.). Decision Trees in Machine Learning: Two Types (+ Examples). [online] Available at: <https://www.coursera.org/articles/decision-tree-machine-learning>.
3. Simplilearn.com. (n.d.). What is Neural Network: Overview, Applications, and Advantages. [online] Available at: <https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-neural-network>.
4. Brownlee, J. (2021). A Gentle Introduction to Multiple-Model Machine Learning. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/multiple-model-machine-learning/>.
5. Google (2019). Classification: Precision and Recall | Machine Learning Crash Course | Google Developers. [online] Google Developers. Available at: <https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall>.