STARS

University of Central Florida
STARS

Data Science and Data Mining

May 2023

Analysis of Credit Approval by Decision Tree

Amir Alipour Yengejeh
amir.alipouryengejeh@ucf.edu

 Part of the Data Science Commons

Find similar works at: https://stars.library.ucf.edu/data-science-mining

University of Central Florida Libraries http://library.ucf.edu

This Article is brought to you for free and open access by STARS. It has been accepted for inclusion in Data Science and Data Mining by an authorized administrator of STARS. For more information, please contact STARS@ucf.edu.

STARS Citation

Alipour Yengejeh, Amir, "Analysis of Credit Approval by Decision Tree" (2023). *Data Science and Data Mining*. 5.
https://stars.library.ucf.edu/data-science-mining/5

University of Central Florida

STARS
Showcase of Text, Archives, Research & Scholarship

Analysis of Credit Approval by Decision Tree

Amir Alipour Yengejeh
dept. Statistics and Data Science
University of Central Florida
Orlando, United States
amir.alipouryengejeh@ucf.edu

***Abstract*—Nowadays, machine learning algorithms are commonly used by the financial institutions or bankers to evaluate the applications' requires for credit card. In this study, we used the decision tree algorithm to predict credit card approval based on the other associated features applicants like *age*, *employment status*, *Education Level*, etc. Our results shows that the applicants' *Prior Default* and *Debt*, and Employed have more contribution in the credit card approval.**

***Index Terms*—Credit Card, Decision Tree, Machine Learning**

I. Introduction

Today, people all around the world submit the applications for the credit card or loan in the financial institutions or banks. Here, the the personal attributes of applicants , in particular the financial history can heavily influence on the institutions' approval decision. For instance, the characteristics such as age, gender, employment status, income, credit history etc can play major role in receiving grant permission. However, it entails the institutions to use the credit analysis techniques to measure the probability whether a third part will repay a loan on time or not. This analysis is essential in reducing the risk of loss for the bankers. Here, there are two basic risks such that bank rejects the qualified applicants and second is subject to loss through accepting unqualified ones. However, manually analyzing of these applications is not only time-consuming but also prone to be error. Today's, data mining techniques or machine learning algorithms are employed by the institutions to do this task automatically.

In this study, we are interested in applying one of the popular and widely used algorithms, decision tree to build a automated model for predict the credit card approval. Decision tree is considered as a non-parametric classification algorithm in which the population is divided into the segments that are similar the branches and build an inverted tree with a root node, internal nodes, leaf nodes. The algorithm can effectively handle huge and large datasets [1] [2]. In this study, our object is building a decision tree model to predict the credit card approval and determine the importance features in this analysis. To do so, we first split our cleaned data into training and validation sets. Then, we build a decision tree model with based on the parameters when both training and test sets have same values.

II. Data set

In this section, we like to glance on the dataset and available in the University of California Irvine (UCI) [3]. The dataset is a good mixture of categorical and continuous attributes but has few missing values. It contains 690 instances with 16 attributes . However, the feature' names was anonymized by the contributors to protect confidentiality of the data, but we assigned some working names on the variables based on their type. The first fifteen variables are applicant attributes but the last variable is credit card approval status or the target variable. The below table summaries the primary information of the attributes and their type.

TABLE I
The List of Features

ID	Attributes	type
0	Male (feat 1)	Categorical
1	Age (feat 2)	Categorical
2	Debt (feat 3)	Continuous
3	Married (feat 4)	Categorical
4	Bank Customer (feat 5)	Categorical
5	Education Level (feat 6)	Categorical
6	Ethnicity (feat 7)	Categorical
7	Years Employed (feat 8)	Continuous
8	Prior Default (feat 9)	Continuous
9	Employed (feat 10)	Continuous
10	Credit Score (feat 11)	Continuous
11	Drivers License (feat 12)	Categorical
12	Citizen (feat 13)	Categorical
13	Zip Code (feat 14)	Categorical
14	Income (feat 15)	Continuous
15	Approved Status (feat 16)	Categorical

Note that the target variable values are "+" or "-" denote whether the credit card request has been approved or not.

III. Data Exploratory

In this section, we are interested in to visualizing and summarizing the properties or description of our attributes in this study. The figure 1 shows our target attribute (approved status) is almost balance such that 383 (55.5%) out of 690 applications were denied while 307 (44.5%) approved. Also, the figure 2 depicts the correlation among numerical attributes in the study. It turns out that there is a weak correlation among these features.

IV. Methodology

In this section, we briefly review the structure of the decision tree classifier, model building, model performance, and data pre-processing.

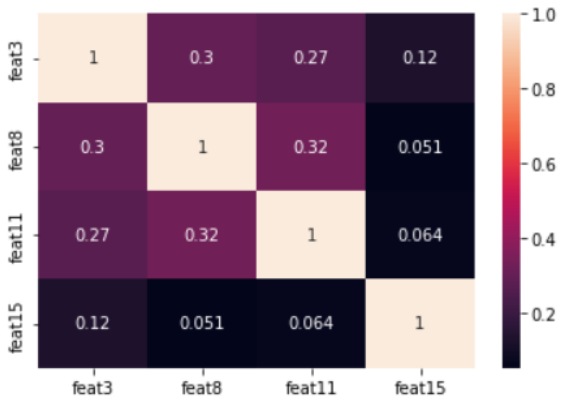

Fig. 1. The correlation matrix of the numerical features.

A. Decision Tree Model Structure

As an supervised learning algorithm, decision tree can be utilized to address the regression and classification problems. but typically preferred for doing classification. Figure 3 illustrates the decision tree. It resembles a tree-structure where the root stand for the attributes, branches for decision rule, and the leaf node for the target variable.

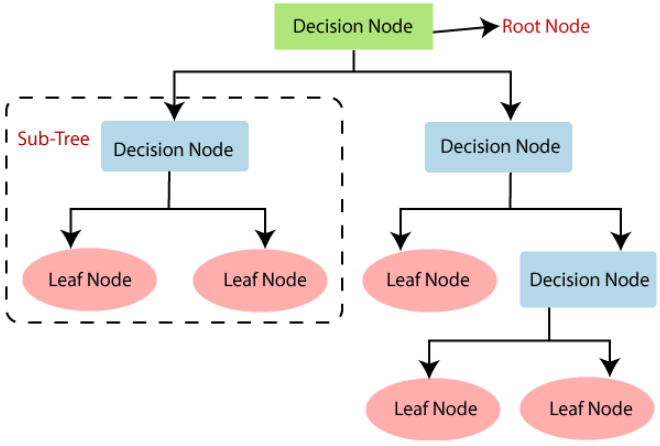

Fig. 2. The general structure of a decision tree: (source: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm)

In decision tree, the goal is splitting the feature space into smallest regions. This process can be continue so much so that there is not any more role to apply on the data points left and a class is s assigned to all data points at each leaf. The *purity of the resulting(child) node* is defined as a *loss function* in decision tree to evaluate each split. In other words, the distribution of a class before and after split is compared by the loss function [4]. In this regard, *Gini Impurity* and *Entropy*.

- **Gini Impurity** measured the spread or variance across the various classes [3].

$$Gini(Node) = \sum_{k=1}^{c} p_k(1 -_p k)$$

where $p_k = \frac{\#observations\ in\ class\ k}{\#observations\ in\ node.}$

- **Entropy** measures inner chaos of the node.

$$Entropy(Node) = -\sum_{k=1}^{c} p_k log(p_k).$$

where the split is locally *local optimal* if the Entropy of child node is lower than the Entropy of the parent one.

B. Building Model

To develop a decision tree that can predict our outcome variable, we need to determine the hyper-parameters like *criterion*, *max_depth*, *min_samples_leaf*, and *min_samples_split*. There are some techniques to find the best parameters to build the model. In this study, we fixed size depth to the intersection of both training and test accuracies.

C. Model Performance

To evaluate the performance our model, we use the below metrics.

- **Confusion Matrix** is a tabular representation of ground-truth (True Values) versus the predicted values by the model. The figure 4 shows the scheme of the confusion matrix.

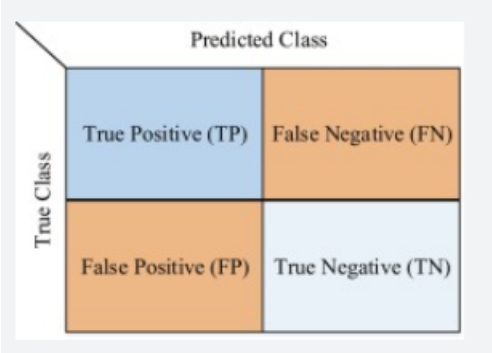

Fig. 3. The diagram of the confusion matrix
source: https://www.sciencedirect.com/topics/engineering/confusion-matrix

- **Model Accuracy** is a ratio of the number of the classes predicted over the total observations:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

- **Precision** is a fraction of true positive and predicted positives,

$$Precision = \frac{TP}{TP + FP}.$$

- **Sensitivity (recall)** is a ratio of true positive to the all true positives,

$$F1Score = \frac{TP}{TP + TN}.$$

- **F1-Score** is harmonic mean of both precision and sensitivity. It is recommended when the outcome variable is imbalance.

$$Sensititivity = \frac{2}{\frac{1}{precision} + \frac{1}{sensitivity}}.$$

D. Data Preparation

As mentioned, the dataset is a mix of categorical and continuous attributes. Before performing the model, the data needs to prepare as follows,

1) The data contains some symbols like "?" and missing values. To handle these entries we first converted the "?" to "NA" and then impute the missing values with the other observations.
2) The data needs to be numerical type before feeding into model. Since the dataset in this study contains some categorical variables, it is required to be converted into the numerical ones. In this regard, all categorical attributes were encoded into the numerical entries.
3) Splitting Data: To build the decision tree model and then validate the performance of which the dataset split into training and test sets. In this regard, the attributes and the target variable split into (70%) and (30%) training and test sets.

V. Results

In this section, we like to discuss about the results of learning decision tree algorithm on our dataset to predict the approved status. In the first step, we learn the model for range number of depths (1 t0 10). The figure 5 shows that the accuracy curves of both training and test sets intersect one another for *max_depth=3*.

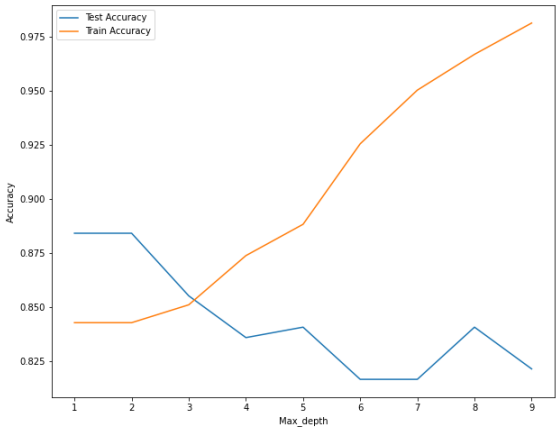

Fig. 4. The train vs test accuracy

Th the model develop with with hyper-parameters, that is *criterion=gini* and *max_depth=3*. Figure 6 represents the the diagram of the decision tree with these parameters. As you can see the tree starts with Prior Default (feat9) as the root node and then separates two regions (branches). For the left branch, the parent node is Debt (feat3) and the right one is Credit Score (feat 11). The figure 7 depicts the confusion matrix of the algorithm. Even though we can see some miss-classifications, most data point classified correctly. Also, the table 2 summarizes the values obtained for different evaluation metrics for both training test sets. As you can see, the accuracy of the model for test set is 0.86 and very close to the training one. Note that the other metric scores are related to the approved class.

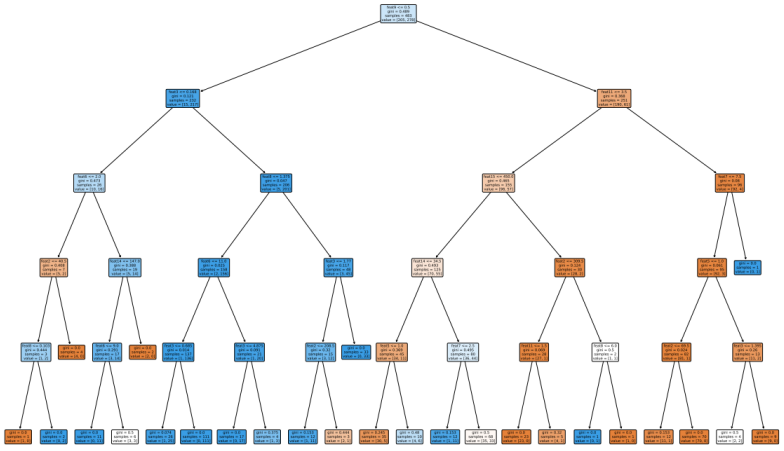

Fig. 5. The diagram of the decision tree

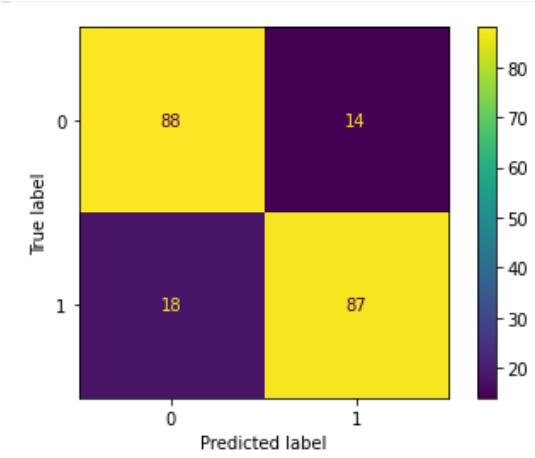

Fig. 6. The confusion matrix of the initial decision tree

TABLE II
The Evaluation Metrics

Data type	Accuracy	Precision	Sensitivity	F1 Score
Train	0.85	0.76	0.95	0.84
Test	0.86	0.83	0.89	0.86

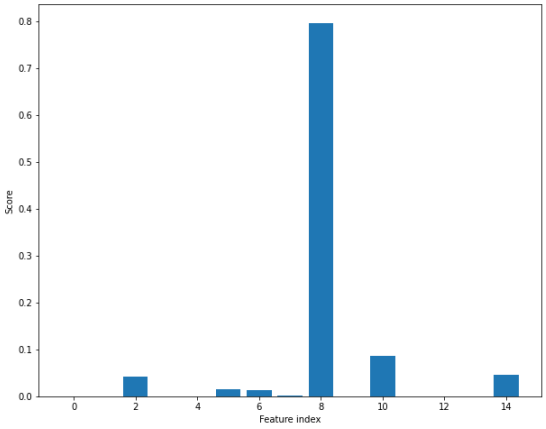

Fig. 7. The importance features

Figure 8 represents the importance score assigned to the attributes by the model. A you can see **Prior Default** of the applicants enjoys the largest score of importance.

VI. Conclusion

In this study, we used the machine learning algorithm to construct an automate credit card approval model based on the information of the applicants. In this regard, we employed a tree-based model called decision tree in which the data split to nodes and leafs based on the related rules. To build such model, we did first some data preparations such as imputing the missing values, and encoding the labels. We fitted the decision tree model by assigning the hyper-parameters like max_depth=3. Th performance of the model in the test set is 0.86 and found the **Prior Default**, **Credit Score**, and **Debt** as importance features respectively to predict the credit card approval. In conclusion, the decision tree shows the enough good performance in predicting the credit card approval, but it would be interested in comparing it with other machine learning models like Xgboost, Logistic regression, etc in the future works.

References

[1] Yan-Yan Song, Ying Lu.(2015) "*Decision tree methods: applications for classification and prediction*". National Institute of Health.

[2] Agbemade, Emil.(2023) "*Predicting Heart Disease using Tree-based Model*."stars.library.ucf.edu

[3] Credit Approval. UCI Machine Learning Repository.

[4] P. Tan, M. Steinbach, and V. Kumar. (2005) *"Introduction to Data Mining"*. Addison Wesley