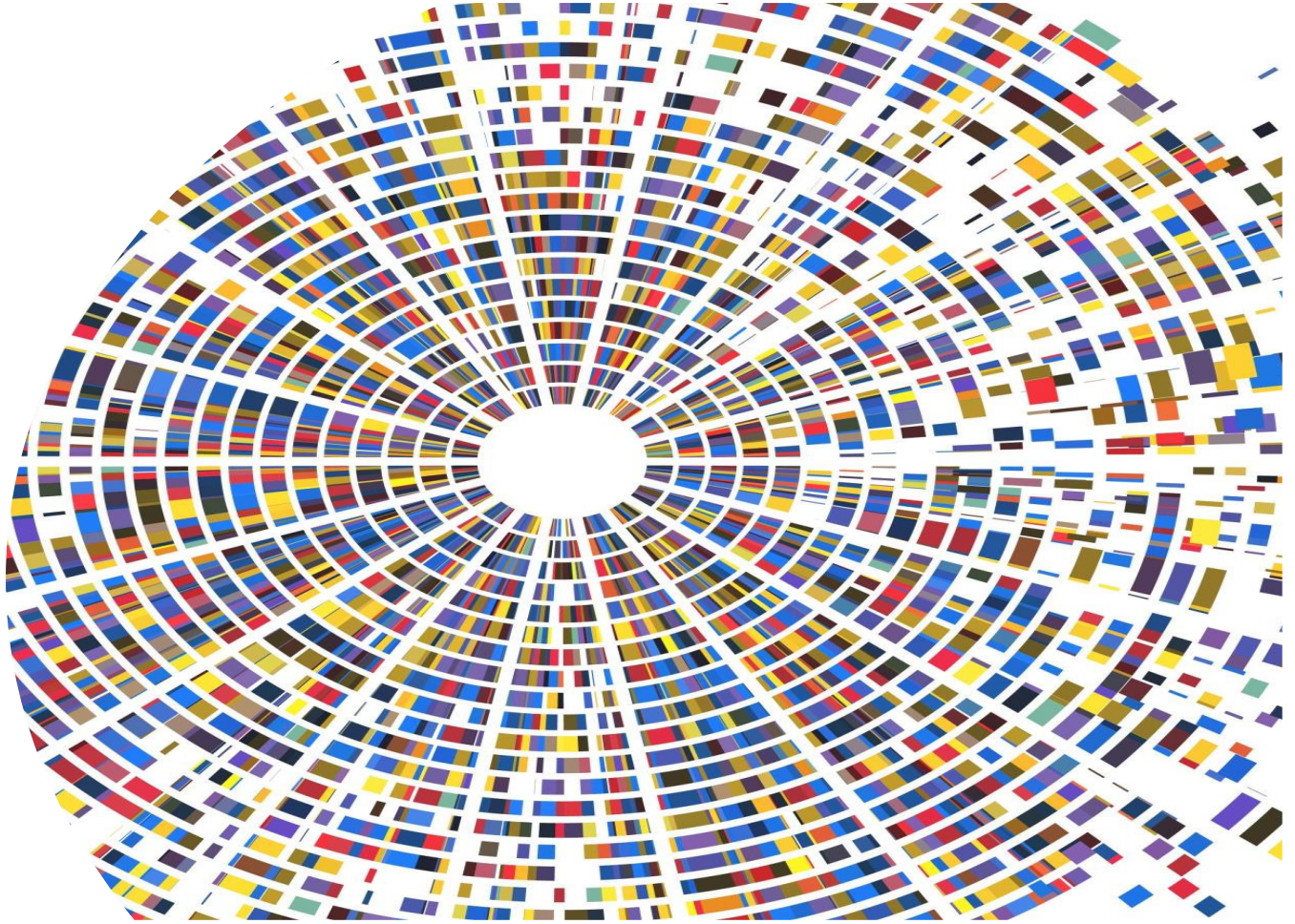


A Targeted Marketing Campaign: Data Analytics Problem



Adesunkanmi Afolabi

Professor: Dr. Liang Chen
Comprehensive Exam Fall 2020
West Texas A & M University
November 18, 2020.

Table of Contents

Cover Page	1
Table of Contents.....	2
Executive Summary.....	3
1. Business Understanding.....	4
a. Background information.....	4
b. Public Reports.....	4
c. Business Problem.....	4-5
d. Brief Action Plan.....	5
2. Data Understanding.....	5
a. Attributes and Records Included in the dataset	5
b. Attributes Meaning	5
c. The Target Attributes and Relationship.....	6
d. The Data Type of Each Attribute.....	6
3. Data Preparation.....	6
a. Data Quality Problems	7
b. Redundant Attributes Treatment and Justifications.....	7-8
c. Visualization of Attributes	8
d. Exploration of the Attributes' Relationship.....	8-9
4. Modeling.....	9
a. Decision Tree and Logistic Regression	9-11
b. Different Data Preparation for Logistic Regression.....	11
c. Advantages and Disadvantages of each Model.....	11-12
d. The Conclusion from the Models.....	12
5. Prediction and Deployment.....	12
a. The Three Models Application to the Prediction dataset.....	12-13
b. Customers Predicted to “Yes” by each model.....	13
c. Prediction Results Comparison.....	13
d. Top 20 Customers for each model.....	13
e. The Highest Expected Return Model.....	13-14
f. Conceptual Automation for Data Mining.....	14
g. Security Concerns for Data Mining on Intranet and Solution.....	15
6. Reflection and Discussion.....	15
a. Problems Encountered and Solution.....	15
b. Limitations and Assumptions	15-16
c. Important Factors Not Considered	16
i. Payback Ability by the Borrowing Customers.....	16
ii. What Happens if a Customer cannot Payback?	16
d. Areas for Future Improvement.....	16
References.....	17
Appendices.....	18-24

Executive Summary

Goal and Motivation

It takes more effort, time, and money to win new customers than to retain the existing ones. This has been universally accepted wisdom in business and marketing. Consequently, the onus has been on every business organization to devise a means to cross-sell products to their existing customers in order to increase their share of wallet of their loyal customers. Failure to do this will naturally make the customers find an alternative business elsewhere which often than not leads to churn in favor of a company that offers diversified products and makes themselves a one-stop-shop. Success in doing this will make the organization keep its customers. In this project, we devise a marketing campaign for a targeted population of a bank's customers.

A target market is defined as "a group of customers within a business's serviceable available market at which a business aims its marketing efforts and resources. A target market is a subset of the total market for a product or service." (*Target Market*, 2020) The goal is to predict the likelihood that 20 out of the 200 targeted customers will likely accept a loan offer. The prediction will be established through data mining models and based on the given attributes.

Method and Data

This project is on the marketing analytics domain and to be able to achieve the objective, secondary data supplied by the professor is utilized. The two datasets are posted on the West Texas A&M University web class blackboard. On the original datasets, data visualization and descriptive analytics were conducted. This leads to the next step; the data was prepared by removing variables that could impact the outcome of the prediction. The predictive analytics component of the project was completed by using the training dataset with a predictive dataset to create two models of a decision tree and one logistic regression model on RapidMiner while the manual identification of the outcome of the data analysis was done using Microsoft excel.

Key Findings and Recommendation.

Using Decision Tree (a): 85 are predicted to respond yes. Using Decision Tree (b): 77 are predicted to respond yes. Using Logistic Regression: 84 customers are predicted to respond yes. I work with the probability of 0.71, 0.8, and 0.42 in each respective model. The expected total return for each model Decision Tree (a), Decision Tree (b), and Using Logistic Regression are: \$142,000.00, \$156,000.00, and \$79,968.00 respectively. Considering model Decision Tree (b) gives the maximum highest return, this model is recommended for adoption for the planned target marketing campaign.

Conclusion

In conclusion, running a productive marketing campaign is the desire of every organization. Running a barren campaign is a major concern. In running a successful campaign, many unexpected factors can influence the decision of customers who have responded "Yes" to a previous campaign and have responded "Yes" to the current campaign. Such inherent factors are not the object of consideration in this analytics project. The output of this analysis can be utilized in any financial institution or marketing department of a company who desires to run a targeted campaign. The existence of historical behavior data of customers helps predict the future behaviors of customers' purchase habits.

1. Business Understanding

a. Background information

b. Public Reports

It is the desire of every organization to grow in profit every year. This ensures they continue in business as a going concern. This makes it indispensable for an organization to hold strategy sessions yearly, quarterly, and some monthly to formulate policy and implementation procedures that will lead to customer acquisition, retention, and profit maximization. To achieve this lofty aim, organizations research their markets and customer patronage that makes it possible for them to create market segmentation. Market segmentation is defined as “a process of dividing a heterogeneous market into relatively more homogenous segments based on certain parameters like geographic, demographic, psychographic, and behavioural” (Wikipedia contributors, 2020). Because one size does not fit all, there is wisdom for market segmentation. It allows for stronger marketing messages, targeted digital advertising, development of effective strategies, better response rates, and lower acquisition costs, attracting the right customers, increasing brand loyalty, branding differentiation from other competitors, identifying niche markets, staying on message, driving growth, enhanced profits, and ideal product development. In fact, according to a study by Bain & Company, “81% of executives found that segmentation was crucial for growing profits. Bain also found that organizations with great market segmentation strategies enjoyed a 10% higher profit than companies whose segmentation wasn’t as effective over a 5-year period” (Market Segmentation: Everything to Know in 2020 //, 2020). All these points to a fact; market segmentation is good and effective for a business to be successful.

c. Business Problem

This analytics project is premised on bank profitability through the effective use of market segmentation. This will allow the bank to increase its market share and the share of the wallet of the customers. The customers have been identified. The propensity of each customer to respond positively to the marketing campaign is the basis of this project. While customer loyalty plays a decisive role in customer identification, even among loyal customers, there is the need to know and understand those who might be interested in responding to a specific loan the bank is offering. This is the basis of the target campaign and the whole essence of this project and therefore the problem it seeks to solve.

The importance of a project like this can be better appreciated when we consider the opinion of Kaplan and Norton (1992) “Managers must identify the customer and market segments in which the company competes and clarify the appropriate measures of performance in these targeted segments. Outcome measures, such as customer satisfaction, customer retention, new customer acquisition, customer profitability, and market share, must be linked to the targeted customer segments in which the business anticipates its greatest potential for growth and profitability” Kaplan and Norton (1992).

According to an outbound agency that provides statistics on organizations and customer acquisition, acquiring a new customer costs anywhere between five times to twenty-five times more than retaining an existing customer. More so, increasing customer retention by 5% can increase profits from 25-95% while the “success rate of selling to a customer you already have is 60-70%, while the success rate of selling to a new customer is 5-20%” (Landis, 2020). It makes logical sense for a smart organization to focus more energy on customer retention by cross-selling products that can increase its share of the customers’ wallet than devoting all energy and attention to acquiring the new ones. The agency goes further to say: “U.S. companies lose \$136.8

billion per year due to avoidable consumer switching” (Landis, 2020). Needless to say, the need to keep the existing customers from escaping cannot be overemphasized. Importantly, this project is a requirement for the final comprehensive examination required for the graduate school of Masters in Computer Information Systems and Business Analytics at West Texas A&M University.

d. Brief Action Plan

The goal of this project is to predict 20 customers who will respond positively to the marketing campaign the bank is embarking on using the existing data based on the customers' historical behaviors. By analyzing the datasets that contain hundreds of records through data analysis in tools such as Excel and RapidMiner, it is possible to unveil the features of the customer in the dataset, predicting the 20 individuals with a high probability of saying “Yes” thereby accepting the loan for the success of the marketing campaign.

To be able to fulfill the goal of this project, the given dataset will be subjected to the data analytical tools: Microsoft excel, RapidMiner. Once the data is prepared, the training dataset will be used to create a decision tree and logistic regression. This will help to determine the best performing model on the prediction data set to predict the likelihood that a given individual in the prediction dataset will accept the loan.

2. Data Understanding

a. Attributes and Records Included in the dataset

To process this dataset, there are 10 attributes included in both the training and predictive datasets. These amount to 7,700 records altogether. The train data contains 10 attributes and 5,900 records while the predictive dataset contains the same 10 identical attributes and 1,800 records because the response attributes values are empty but depicted with “?” that occupied the blank spaces. The cleaned data used for this analysis is different from the original dataset that contained some missing and almost impossible values, outliers.

b. Attribute Meanings

The original dataset has 12 attributes before it was pruned down to 10. Income and Children attributes have to go as discussed under the Data Preparation. The meaning of each attribute is given as follows: (1) **ID**: This provides unique to identify each customer. (2) **Age**: This represents a customer's age (years old) when receiving the loan offer. (3) **Sex**: This is to indicate the customer's gender which can either male, female, or other. (4) **Region**: This attribute represents the region in which a customer is living when receiving the loan offer which can either be Inner City, Rural, Suburban, or Town. (5) **Income**: This attribute indicates a customer's annual taxable income when receiving the loan offer. (6) **Married**: This helps to understand the marital status of each customer whether a customer is married or not. (7) **Children**: This assists us to know the number of children that a customer has when receiving the loan offer. (8) **Car**: The car attribute helps understand if a customer owns a car or not when receiving the loan offer. (9) **Save act**: This is a categorical variable that helps know whether a customer has a saving account or not when receiving the loan offer. (10) **Current act**: This is to determine if a customer has a current loan account with the bank or not at the time of receiving the loan offer. (11) **Mortgage**: This is to determine if the customer has a mortgage or not when receiving the loan offer from the bank. (12) **Response**: This is the attribute that helps us determine if the customer response to the loan offer is either yes or no.

Table 1: Variable definition and measurement

Variable	Definition	Measurement
Id	To identify each customer	Unique ID
Age	A customer's age (years old) when receiving the loan offer	Number in Years
Sex	A customer's gender	Male / Female / Other
Region	The region in which a customer is living when receiving the loan offer	Inner City, Rural, Suburban, Town
Income	A customer's annual taxable income when receiving the loan offer	Amount in numbers
Married	Is this customer married when receiving the loan offer?	Yes / No
Children	The number of children that a customer has when receiving the loan offer	Count in numbers
Car	Does this customer own a car when receiving the loan offer?	Yes / No
Save_act	Does this customer have a saving account when receiving the loan offer?	Yes / No
Current_act	Does this customer have a current loan account with the bank?	Yes / No
Mortgage	Does this customer have a mortgage when receiving the loan offer?	Yes / No
Response	Does this customer respond to the loan offer	Yes / No

c. The Target Attributes and Relationship

“Response” is the target variable here and therefore the dependent variable while all other variables: ID, Age, Sex, Region, Married, Car, Save_act, Current_act, and Mortgage are treated as independent variables.

d. The Data Type of Each Attribute

Table 2: Attribute Data Type

Attribute Name	Description	Data Type
Id	To identify each customer	Numerical
Age	A customer's age (years old) when receiving the loan offer	Discrete / Integer
Sex	A customer's gender	Categorical/Binomial
Region	The region in which a customer is living when receiving the loan offer	Categorical/Polynomial
Income	A customer's annual taxable income when receiving the loan offer	Numerical/Continuous
Married	Is this customer married when receiving the loan offer?	Categorical/Binomial
Children	The number of children that a customer has when receiving the loan offer	Discrete/integer
Car	Does this customer own a car when receiving the loan offer?	Categorical/Binomial
Save_act	Does this customer have a saving account when receiving the loan offer?	Categorical/Binomial
Current_act	Does this customer have a current loan account with the bank?	Categorical/Binomial
Mortgage	Does this customer have a mortgage when receiving the loan offer?	Categorical/Binomial
Response	Does this customer respond to the loan offer	Categorical/Binomial

3. Data Preparation

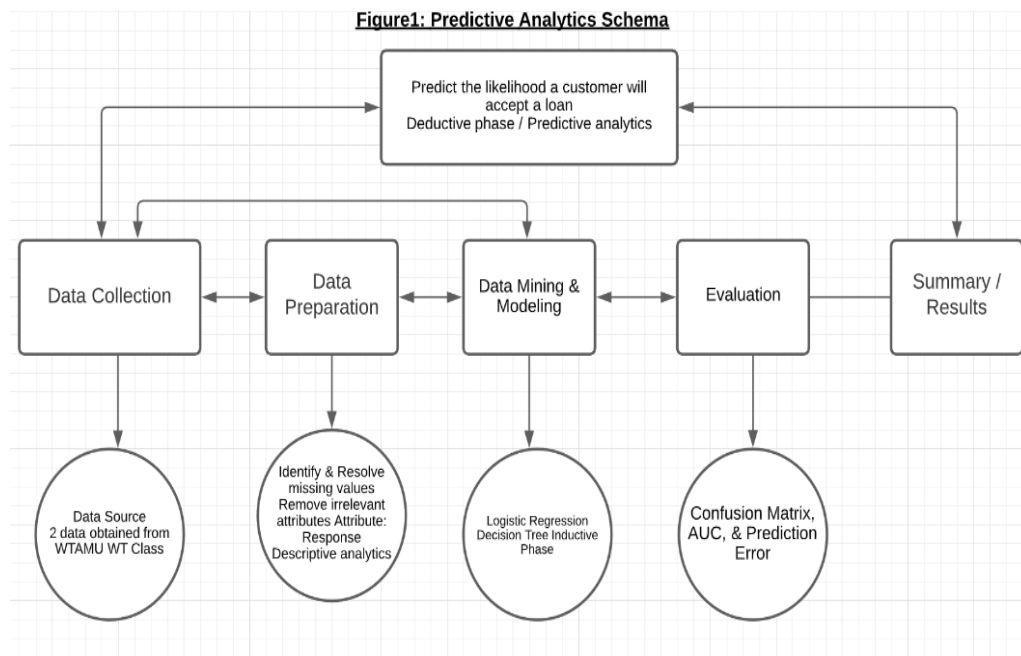
Data preparation takes time. It takes 80% of the time for data scientists when working on a project (Press 2016). This is expected as the preparation must be right in order to have reliable data analysis. The data was cleaned to get the correct outcome. The project method is referred to as a “classic machine learning procedure” that utilizes the scientific paradigm of induction and deduction (University of Eastern Finland, n.d.). In the inductive phase, machine learning models on secondary datasets retrieved from the WTAMU blackboard are trained.

The inductive phase helps create models that learn general rules from our dataset; general rules which accept “positive examples” and reject “negative examples” via a process called supervised machine learning. The structure of the dataset makes the decision tree the ideal model. In the deductive phase (predictive phase), the best performing model from the inductive phase is used to make predictions using new data (second dataset). This schema includes data collection, data preparation, model building, model evaluation, and model deployment to make predictions as depicted in Figure 1.

The data analysis was conducted using descriptive analytics techniques. This gives room for data visualization for a graphical understanding of different attributes available in the dataset. The target variable is a response to know those who answered “yes” to the previous marketing campaign and those who answered “no” to them. The predictors or independent variables are age, sex, region, income, married, save_act, and current_act. car and mortgage are controlled variables. The supervised machine learning method enables the understanding and identification of the attributes that are important predictors of the target variable “response.”

This uses relevant attributes to predict the probability that a customer will accept the loan. The project is marketing analytics. Descriptive analytics (D) is used to visualize and better understand the data while predictive analytics (P) is used to make a prediction. The outcome will bring Innovation (I) and Agility (A) to bank loan marketers, which will increase their Productivity (P). By developing a predictive model, bankers and other financial service organizations can expeditiously and accurately predict which customers will accept a loan offering with alacrity. This will allow them to quickly make such loans available to the customers

who will not refuse. It eliminates wasting spending on campaigns that would have been directed to unwilling customers who will ultimately reject such loans.



a. Data Quality Problems

b. Redundant Attributes Treatment and Justifications

The original dataset was supposed to contain 9,412 records. The training dataset is supposed to be 7,212 with 12 attributes while the predictive dataset is supposed to be 2,200 on the face of it. In reality, this is not the case. The original predictive dataset has 200 missing values. This is because the “Response” is empty only occupied with “?” in place of a null value. The training dataset is the most capricious. It contains 425 missing values out of the 7,212 records. These amount to approximately 6% of the records. This is besides the duplicate record. This problem needed to be addressed in order to balance the data and have clean data for the analysis.

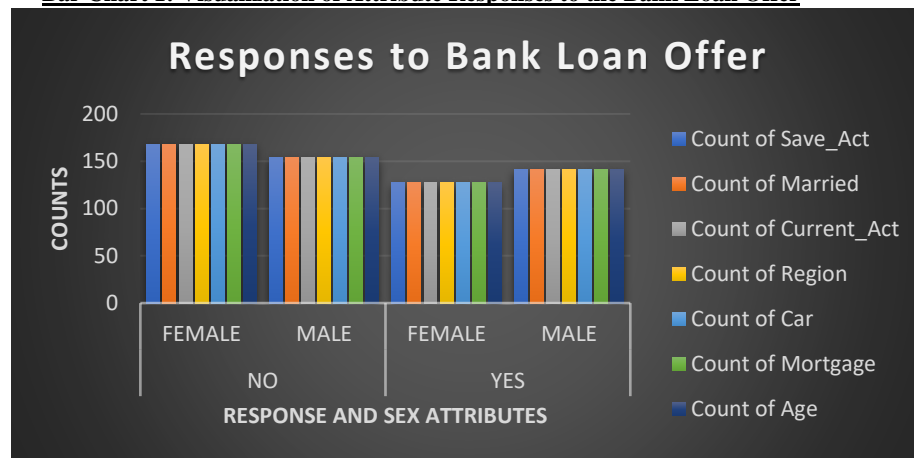
The training dataset data quality problems were resolved in the following order. A duplicate record of customer Identity ID12101 was discovered and deleted. The age record has some outrageous values for “Age” such as 999 for 3 different customers, 420 years for 1 customer, and 158 years for one customer. A quick check on the Guinness World Records reveals the verified oldest living person, Kane Tanaka, is 117 years, 336 days and she lives in Japan, not in America

(Punt, 2020). This informs the decision to delete the 5 records under “Age” attributes. Not only do they stand out as extreme values but are also capable of jeopardizing the data integrity as outliers. More so, there are 8 missing values under the Age attribute. To solve this problem, the mean of the values of the Age attribute is calculated after the outrageous age numbers were deleted. This comes to 41.12. Hence, 41years is entered in for each of the 8 missing values. Additionally, there are 5 missing values under the “Car” attribute. This is less than 1% of the record. Since the Car attribute contains a categorical binomial response of either Yes or No, it is considered unwise to apply neither the mean nor mode. Consequently, the 5 records were deleted to preserve the integrity of the dataset. The “Children” attribute was deleted in the training dataset because it is absent in the predictive dataset attributes. Finally, on the training dataset, the “Income” attribute contains 412 missing values out of possible 600 values. This is about 69% of the Income record in the training dataset. After a lot of ambivalence, the Income attribute was deleted from both the training dataset and predictive dataset. This is because neither, mean, mode, nor record deletion could salvage the “Income” attribute with about 70% missing values. After all these, we now have decent data that we can work on with peace of mind.

c. Visualization of Attributes and

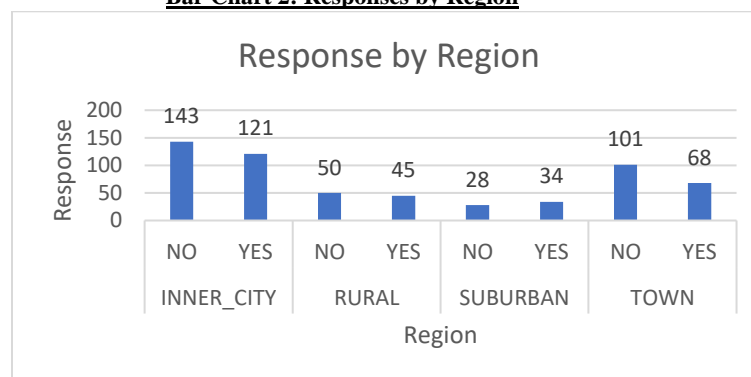
d. Exploration of the Attributes’ Relationship

Bar Chart 1: Visualization of Attribute Responses to the Bank Loan Offer



The No Response is more pronounced among the females than their counterparts. This is eloquently visible among all the attributes: Save_Act, Married, Current_Act, Region, Car, Mortgage, and Age. Males respond more positively to bank loan offer than the Female. Hence, the highest positive response to this bank will come from the males.

Bar Chart 2: Responses by Region



Even though the highest number of No responses comes from the Inner_City, interestingly, the response that also carries highest number of Yes also falls into the same Inner_City region.

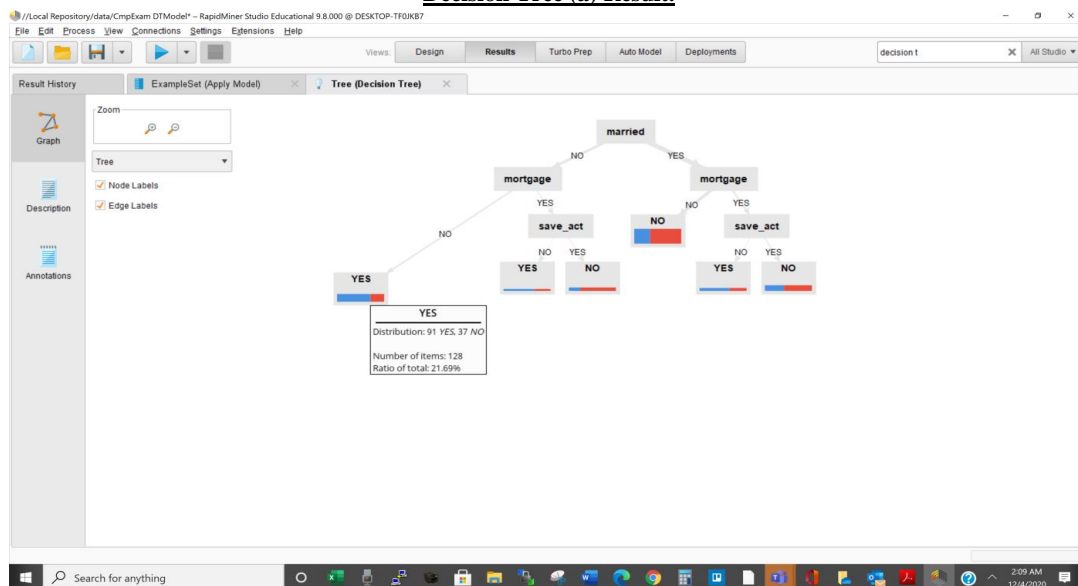
4. Modeling

a. Decision Tree and Logistic Regression

The importance of the decision tree is fully experienced in this project. This is because when all means of solving the question came to halt. It was the decision tree that assisted to remove the impasse and get a meaningful solution. No wonder, it is said: “Decision trees provide an effective method of Decision Making because they: Clearly, layout the problem so that all options can be challenged. Allow us to analyze fully the possible consequences of a decision. Provide a framework to quantify the values of outcomes and the probabilities of achieving them” (Decision Tree Analysis: Choosing by Projecting “Expected Outcomes,” 2020).

This project would have been impossible without the use of a decision tree on RapidMiner. Using Minimal Gain of 0.01, Minimal Size for Split 10, and Minimal Leaf Size of 20, the result of the decision tree on RapidMiner is as detailed below. The result reveals 91 out of 128 are predicted to accept the offer. This is a probability of 0.71. The ideal person is not married and has no mortgage. Hence, the top 20 customers come from this group. Most importantly, they should be living in the Inner City for the maximum expected return.

Decision Tree (a) Result.



We can see that married is the root node, but it also appears to be a split node at some other parts of the tree. According to the decision tree, for customers who are not married, have no mortgage, 91 respond yes, 37 no. Customers that are not married, have a mortgage but no savings account, 15 respond yes and 8 no. Customers that are not married, have a mortgage and a saving account, 12 respond yes and 38 no. Conversely, customers that are married, have no mortgage 87 respond yes, 170 no. For customers that are married and have a mortgage but have no savings account, 26 respond yes, 15 no. Customers that are married and have a mortgage but have a savings account, 37 respond yes, 54 no.

Decision Tree (b)

Changing the data a bit by using Minimal Gain of 0.001, Minimal Size for Split 4, and Minimal Leaf Size of 2, the result of the decision tree on RapidMiner is as detailed below. The result

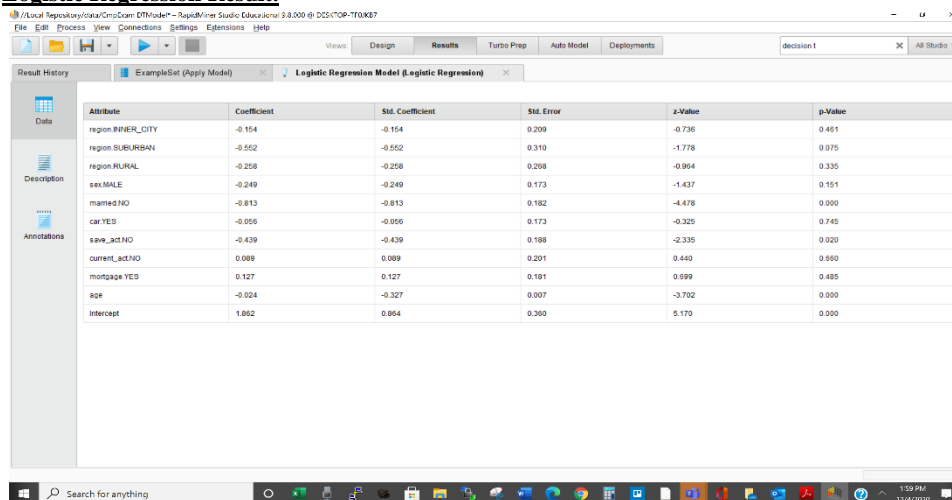
car. 1 yes, 1 no. With car none yes, 3 no. less than and equals 64 without a savings account, age greater than 59, 3 yes one no. Less than and equals 59 lives in Inner-city, 11 yes, 5 no, Lives in Rural, 3 yes and 1 no. Lives in Suburban, 4 yes and 1 no. Those who live in Town who are female, none yes, 4 no while Male respond same. 3 yes and 3 no. Those who have a savings account, older than 21 years and older than 38 years, and older than 40 years, 24 yes 27 no. Less than but equals 40 years, 4 yes and none no. Less than and equals 38 years 8 yes and 19 no. Less than and equals 21 years, none yes and 2 no. For those who are less than and equals to 19 years, 1 yes and 9 no.

b. Different Data Preparation for Logistic Regression

Logistic Regression

The data for the Logistic Regression was prepared without including the ID attribute. The result on the RapidMiner reveals attributes of customers with no current account and those that have a mortgage are the two attributes that have positive coefficients of 0.089 and 0.127 respectively. Negative coefficients indicate that the event is less likely at that level of the predictor. More so, the age, customers with no savings account are considered statistically significant. The rest are not because the P-Values are greater than 0.05. They are not significantly relevant in predicting the customers who will likely accept the bank loan offer. Hence, in arriving at the top 20 customers, cognizance is taken of the statistically significant attributes: Married.NO, Save_Act.No, and Age. They are not married. They have no saving account.

Logistic Regression Result.



Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
region INNER_CITY	-0.154	-0.154	0.209	-0.736	0.461
region SUBURBAN	-0.552	-0.552	0.319	-1.778	0.075
region RURAL	-0.258	-0.258	0.208	-0.964	0.335
sex MALE	-0.249	-0.249	0.173	-1.437	0.151
married NO	-0.813	-0.813	0.182	-4.478	0.000
car YES	-0.056	-0.056	0.173	-0.325	0.745
save_act NO	-0.439	-0.439	0.188	-2.335	0.020
current_act NO	0.089	0.089	0.201	0.443	0.660
mortgage YES	0.127	0.127	0.181	0.699	0.485
age	-0.024	-0.327	0.007	-3.792	0.000
Intercept	1.062	0.064	0.360	5.170	0.000

c. Advantages and Disadvantages of each Model

Decision Tree (a): This uses Minimal Gain of 0.01, Minimal Size for Split 10, and Minimal Leaf Size of 20. **Decision Tree (a) Advantage:** It gives the quick result to determine the 20 customers we are looking for at a glance. **Decision Tree (a) Disadvantage:** Not as specific as Decision Tree (b). To get the exact result one still needs to drill down on the data.

Decision Tree (b): This uses Minimal Gain of 0.001, Minimal Size for Split 4, and Minimal Leaf Size of 2. **Decision Tree (b) Advantage:** This is the most specific method. It gives a specific result and specific direction to the exact result. **Decision Tree (b) Disadvantage:** It is more difficult to visualize the result at a glance. It is less organized on the RapidMiner.

Logistic Regression Advantage: It avoids confounding effects by analyzing the association of all variables together as it allows us to obtain an odds ratio in the presence of more than one explanatory variable. **Logistic Regression Disadvantage:** It is less specific. It does not

allow us to directly pin-point the 20 customers who will directly accept the loan as we saw in the Decision Tree models. Despite the mass of data modeled, it only gives the probability of acceptance which is vaguer.

d. The Conclusion from the Models

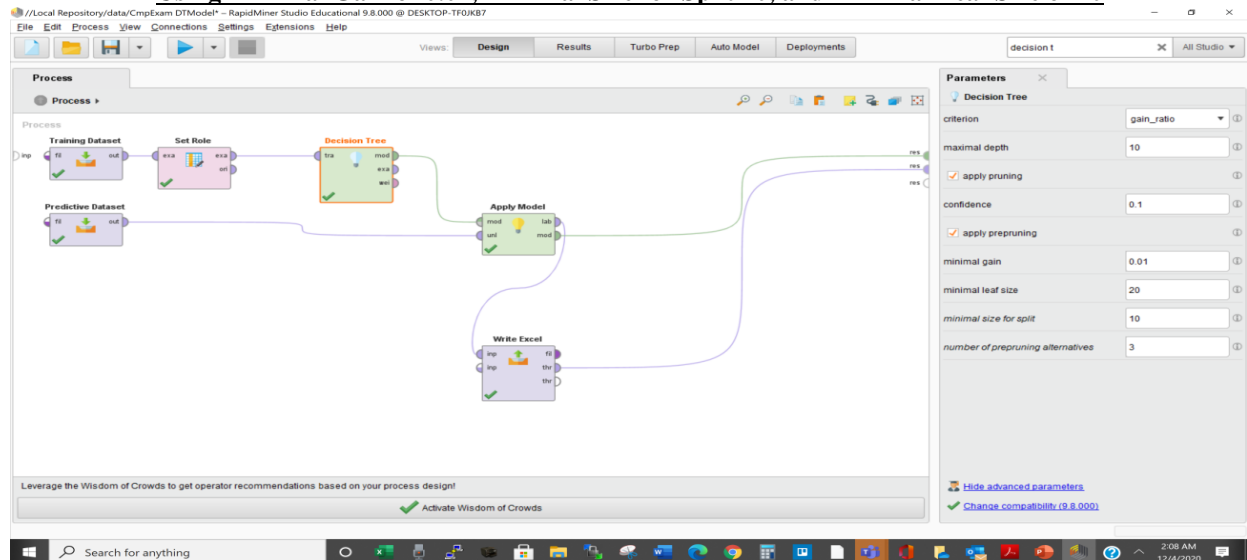
In predicting the result to the expected solution for this project, the Decision tree model appeals to me best. This is because it allows me to be able to easily narrow my result. And distill the needed 20 customers. Also, I will prefer the Decision Tree (b). This is because it is easier to graphically know what the solution is and go for it among the dataset. It is detailed which makes the facts unmistakable and the eventual result not unexpected.

5. Prediction and Deployment

a. The Three Models Application to the Prediction dataset

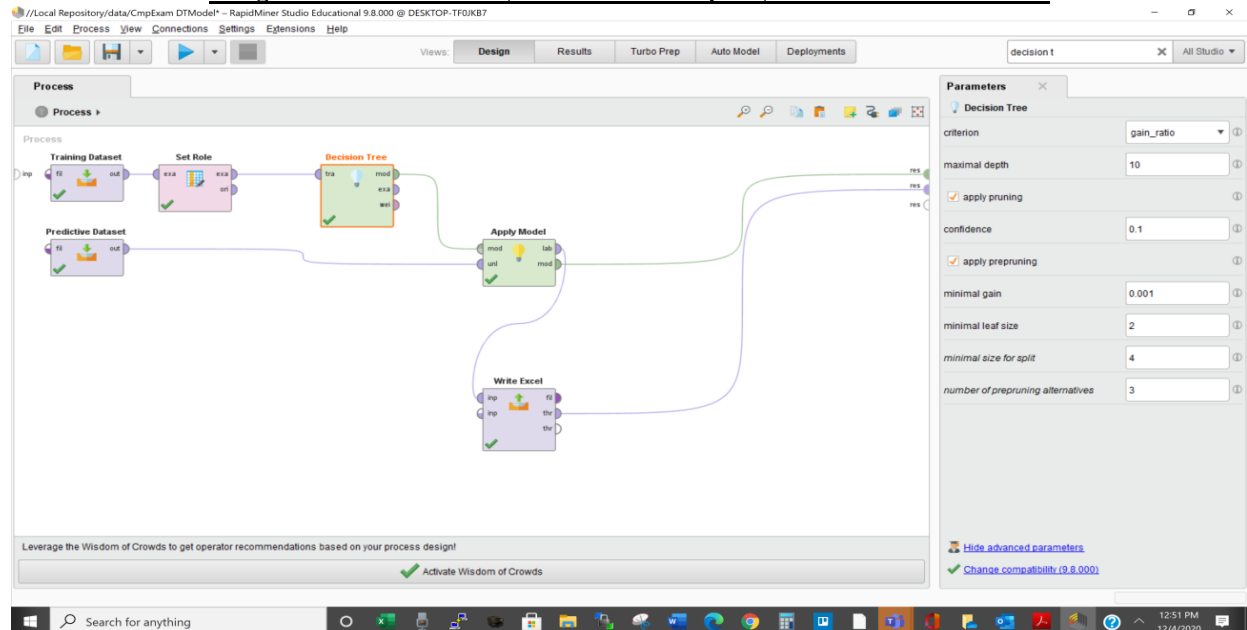
Decision Tree (a)

Using Minimal Gain of 0.01, Minimal Size for Split 10, and Minimal Leaf Size of 20



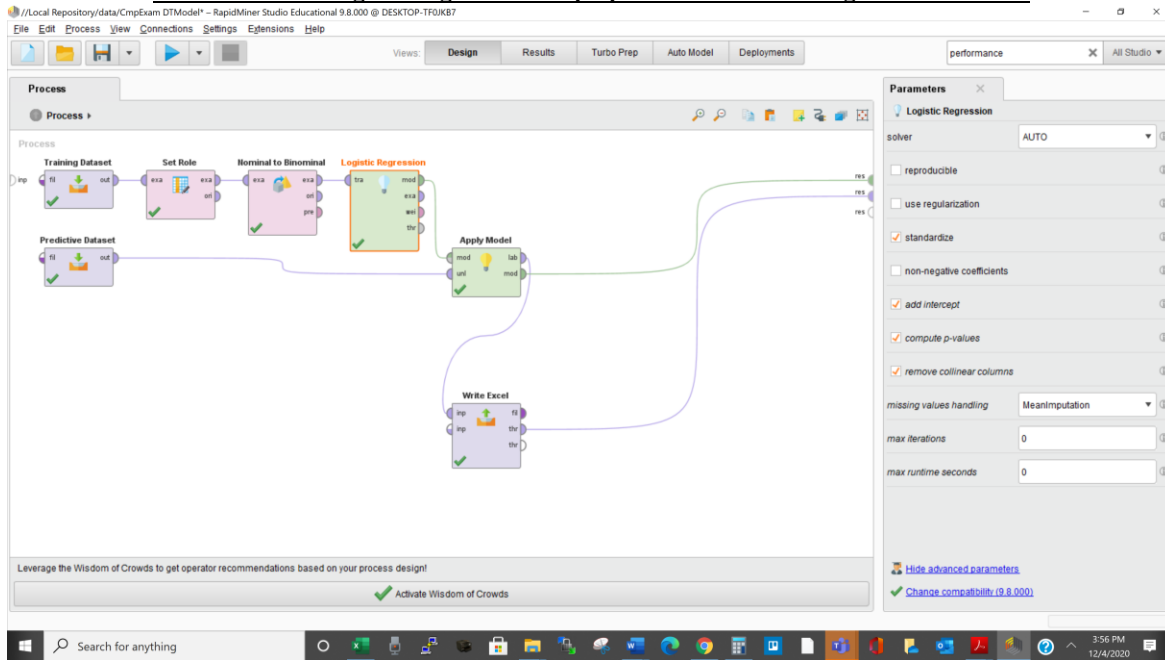
Decision Tree (b)

Using Minimal Gain of 0.001, Minimal Size for Split 4, and Minimal Leaf Size of 2



Logistic Regression

The data for the Logistic Regression was prepared without including the ID attribute.



b. Customers Predicted to “Yes” by each model

- a. Decision Tree (a): 85 Customers
- b. Decision Tree (b): 77 Customers
- c. Logistic Regression: 84 Customers

c. Prediction Results Comparison

The three prediction results are not the same. They are different. The Decision Tree (a) is different from the Decision Tree (b) while Logistic Regression is different from both.

d. Top 20 Customers for each model

Decision Tree (a)					Decision Tree (b)					Logistic Regression			
ID	Region	Fixed Return	Total Return		ID	Region	Fixed Return	Total Return	S/N	ID	Region	Fixed Return	Total Return
ID12709	INNER_CITY	10,000	7,100.00		ID12709	INNER_CITY	10,000	8,000.0	1	ID12712	INNER_CITY	10,000	4,200
ID12730	INNER_CITY	10,000	7,100.00		ID12730	INNER_CITY	10,000	8,000.0	2	ID12746	INNER_CITY	10,000	4,200
ID12746	INNER_CITY	10,000	7,100.00		ID12747	INNER_CITY	10,000	8,000.0	3	ID12748	INNER_CITY	10,000	4,200
ID12747	INNER_CITY	10,000	7,100.00		ID12753	INNER_CITY	10,000	8,000.0	4	ID12759	INNER_CITY	10,000	4,200
ID12748	INNER_CITY	10,000	7,100.00		ID12757	INNER_CITY	10,000	8,000.0	5	ID12760	INNER_CITY	10,000	4,200
ID12753	INNER_CITY	10,000	7,100.00		ID12766	INNER_CITY	10,000	8,000.0	6	ID12771	INNER_CITY	10,000	4,200
ID12757	INNER_CITY	10,000	7,100.00		ID12770	INNER_CITY	10,000	8,000.0	7	ID12773	INNER_CITY	10,000	4,200
ID12759	INNER_CITY	10,000	7,100.00		ID12779	INNER_CITY	10,000	8,000.0	8	ID12783	INNER_CITY	10,000	4,200
ID12766	INNER_CITY	10,000	7,100.00		ID12788	INNER_CITY	10,000	8,000.0	9	ID12810	INNER_CITY	10,000	4,200
ID12770	INNER_CITY	10,000	7,100.00		ID12804	INNER_CITY	10,000	8,000.0	10	ID12814	INNER_CITY	10,000	4,200
ID12771	INNER_CITY	10,000	7,100.00		ID12811	INNER_CITY	10,000	8,000.0	11	ID12817	INNER_CITY	10,000	4,200
ID12773	INNER_CITY	10,000	7,100.00		ID12830	INNER_CITY	10,000	8,000.0	12	ID12832	INNER_CITY	10,000	4,200
ID12779	INNER_CITY	10,000	7,100.00		ID12831	INNER_CITY	10,000	8,000.0	13	ID12852	INNER_CITY	10,000	4,200
ID12788	INNER_CITY	10,000	7,100.00		ID12897	INNER_CITY	10,000	8,000.0	14	ID12856	INNER_CITY	10,000	4,200
ID12804	INNER_CITY	10,000	7,100.00		ID12900	INNER_CITY	10,000	8,000.0	15	ID12741	SUBURBAN	9,000	3,780
ID12811	INNER_CITY	10,000	7,100.00		ID12797	SUBURBAN	9,000	7,200.0	16	ID12756	SUBURBAN	9,000	3,780
ID12814	INNER_CITY	10,000	7,100.00		ID12808	SUBURBAN	9,000	7,200.0	17	ID12758	SUBURBAN	9,000	3,780
ID12817	INNER_CITY	10,000	7,100.00		ID12845	SUBURBAN	9,000	7,200.0	18	ID12840	SUBURBAN	9,000	3,780
ID12830	INNER_CITY	10,000	7,100.00		ID12853	SUBURBAN	9,000	7,200.0	19	ID12720	TOWN	7,200	3,024
ID12831	INNER_CITY	10,000	7,100.00		ID12868	SUBURBAN	9,000	7,200.0	20	ID12739	TOWN	7,200	3,024
Total			\$ 142,000.00		Total			\$ 156,000.00		Total			\$ 79,968.00

- a. Decision Tree (a): computed with 0.71 probability
- b. Decision Tree (b): computed with 0.8 probability
- c. Logistic Regression: computed with 0.42 probability

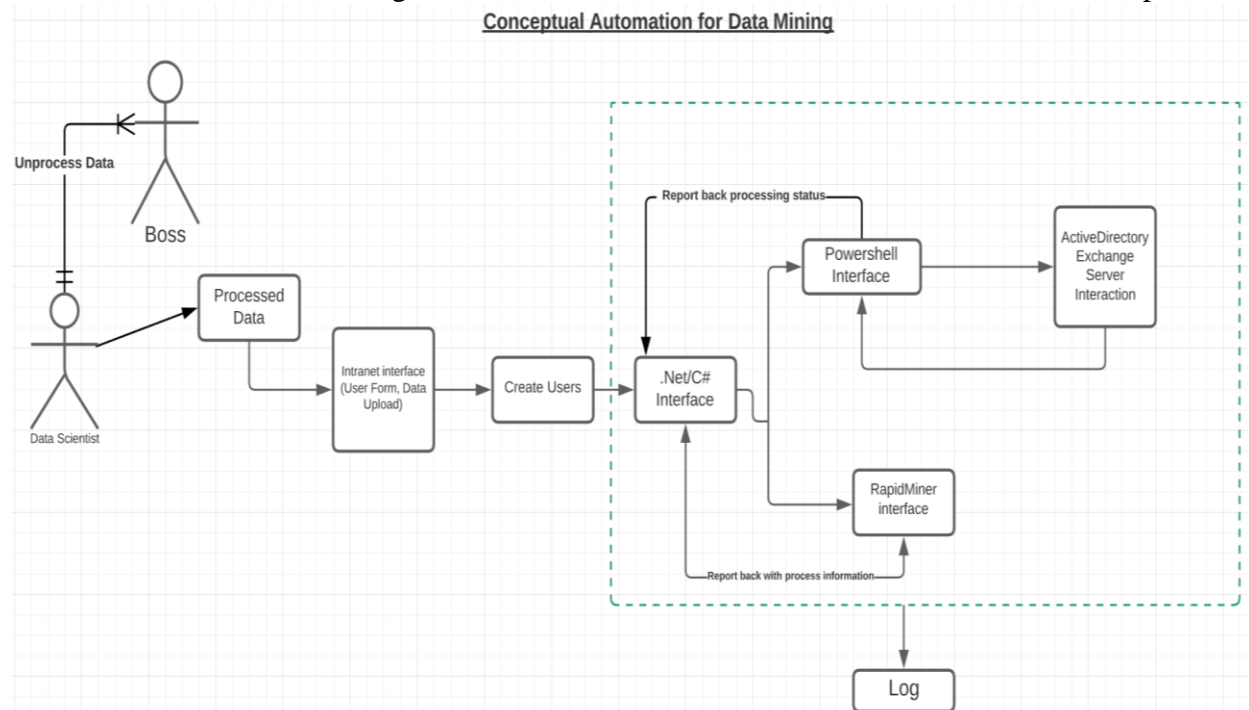
e. The Highest Expected Return Model

The model that will generate the highest return is model b. The Decision Tree model utilizes Minimal Gain of 0.001, Minimal Size for Split 4, and Minimal Leaf Size of 2. Consequently, it is my recommendation that the bank adopts model b to be able to generate the maximum return. These findings support the argument “companies that take a proactive stance in using customers and suppliers as a key source of inspiration, rather than merely monitoring and imitating what competitors are doing, are those that can gain greater rewards in the marketplace and earn a higher market share with better brand awareness in their respective industries” (Dawson & Andriopoulos, 2017, p. 774).

f. Conceptual Automation for Data Mining

This conceptual automation framework is prepared based on the recommendation that only the authorized users will have access to the data and the processed data is already deployed on the RapidMiner software that integrated unto the bank intranet enterprise software server. And that the Bank has FIM: Federated Identity Management.

The boss sends the original dataset to the data scientist via email. The relationship



between the data scientist and the boss is one to one. The data scientist can only have one reporting boss at a time. The relationship between boss and the data scientist is one to many. This means the boss can have many subordinates reporting to him at the same time. The data scientist cleans up the data. He uploads the processed data unto the intranet. From there the software developer or intranet administrator takes over and create the authorized users who by policy should have access to the customers data. They are created on the bank’s database backbend with view rights and editing rights as policy allows for each employee. Once the processed data is uploaded unto the intranet. The process becomes automated. The users can login to the intranet and input through the interface. The RapidMiner offers intelligent data mining and marketing reporting through customizable dashboard. It takes information through the intranet interface and report back with the process information. It feeds information into the interface application from where the PowerShell can give a status reporting after communication with the active directory exchange server on the premise. Active Directory helps organize

company's users, and information going out of the server. The IT admin uses AD to organize the company's complete hierarchy which computers belong to which network to ensure only and only the authorized users have access.

Even if in the near future bank chooses any product offering through target campaign all the user has to do is to login through the bank's intranet and access the RapidMiner. The intended prediction will be displayed through the dashboard on the bank's intranet. The users can then go on to customize the desired report. The log is available to track users and every query run.

g. Security Concerns for Data Mining on Intranet and Solution

(i) **Data Breach through Proliferation of Access to Customers ID**; It is rightly said that when the purpose of anything is not known abuse is inevitable. There are customers ID if it is hosted on the intranet where every employee has uncontrolled access to it, there will be a data breach that exposes confidential, sensitive or protected information to an unauthorized person within the organization. **Solution**: Only the bank staff who are core staff of the bank, not just temporary staff like contractors, should have access to the data mining on the intranet. Secondly, the accessibility should depend on the relevance to the job of the staff. Those who do not need it should not have access to it. Many good practices should be put in place. Employees should be trained in how to avoid being caught by phishing attacks, and how to practice good password protection.

(ii) **Insider Abuse**: Fraud rarely succeeds without internal connivance. Employees can access the information to perpetrate fraud like cloning credit and debit cards and use them to carry out transactions without the knowledge of customers. **Solution**: This can be achieved by monitoring the intranet network with audit and transaction logs. A solution like Liquid Web's custom Malicious Activity Detector (MAD) will also guard against threats from within the organization. If a malicious insider attack is detected, the insider's access privileges should be revoked immediately.

6. Reflection and Discussion

"Predictive analytics are used to determine customer responses or purchases, as well as promote cross-sell opportunities. Predictive models help businesses attract, retain, and grow their most profitable customers. Improving operations. Many companies use predictive models to forecast inventory and manage resources" (Predictive Analytics: What It Is and Why It Matters, 2020). The task of identifying the 20 customers out of the morass of 7,700 is daunting and intensive. The use of predictive analytics helps simplify the process and facilitates the eventual arrival at the expected solution.

a. Problems Encountered and Solution

The data contains a lot of extraneous that would not have facilitated decent prediction. There were 5 outliers under "Age" attributes that ranges between 158 to 999 years of age that were deleted. Cars contained 5 missing values that lead to the deletion of the 5 rows of records in the training dataset. Average was used to resolve the problem of 8 missing values under "Age" while income attribute was deleted in both training dataset and predictive dataset because of 412 missing values in the original training dataset.

b. Limitations and Assumptions

The data analytics project is limited for two reasons. There are 412 missing values in the Income attribute of the train datasets. Income should have been an important attribute to be present

to help understand the earning of the customers who respond positively to campaigns. The attribute “children” is not present in the predictive dataset. The “children” were deleted from the training dataset before the decision tree could be run on the RapidMiner. Again, the presence of the Children attributes would have assisted to know the number of children of customers who respond to loan, whether their number of children is a factor or not. Also, manual sorting on Microsoft excel helps distill the result from the processed files. I wish all are automated. On the conceptual automation, it is assumed that the bank uses RapidMiner for its data mining. And that it has FIM: Federated Identity Management that offers single sign access to a number of applications, including RapidMiner, across various enterprises. More so, it is assumed that the data scientist had deployment the processed data on the RapidMiner before the bank can access it through the web interface.

c. Important Factors Not Considered

- i. Payback Ability by the Borrowing Customers
 - ii. What Happens if a Customer cannot Payback?
- i. Even though these data were processed, and the final answer is gotten, the ability to pay back the loan is not one of the attributes. This is an important factor because, at the end of the day, it is not about customers accepting loans as a result of a direct marketing campaign, their ability to repay the loan after they have been awarded should equally be a factor.
- ii. For a complete data set that can assist quick decision making, there is the need to include the remedy for a failure. Will be covered by insurance or written off by the Federal Deposit Insurance Corporation (FDIC)? Are there specific customers whose bad loans cannot be written off? It will be nice to know.

d. Areas for Future Improvement

I encourage further research into the characteristic of bank customers who easily repay their loans, those who do not pay, and those whose loans cannot be guaranteed by the bank’s insurance. This will help to accurately calculate the expected returns to prevent a situation where the expected return will turn to expected loss after a successful target campaign.

References

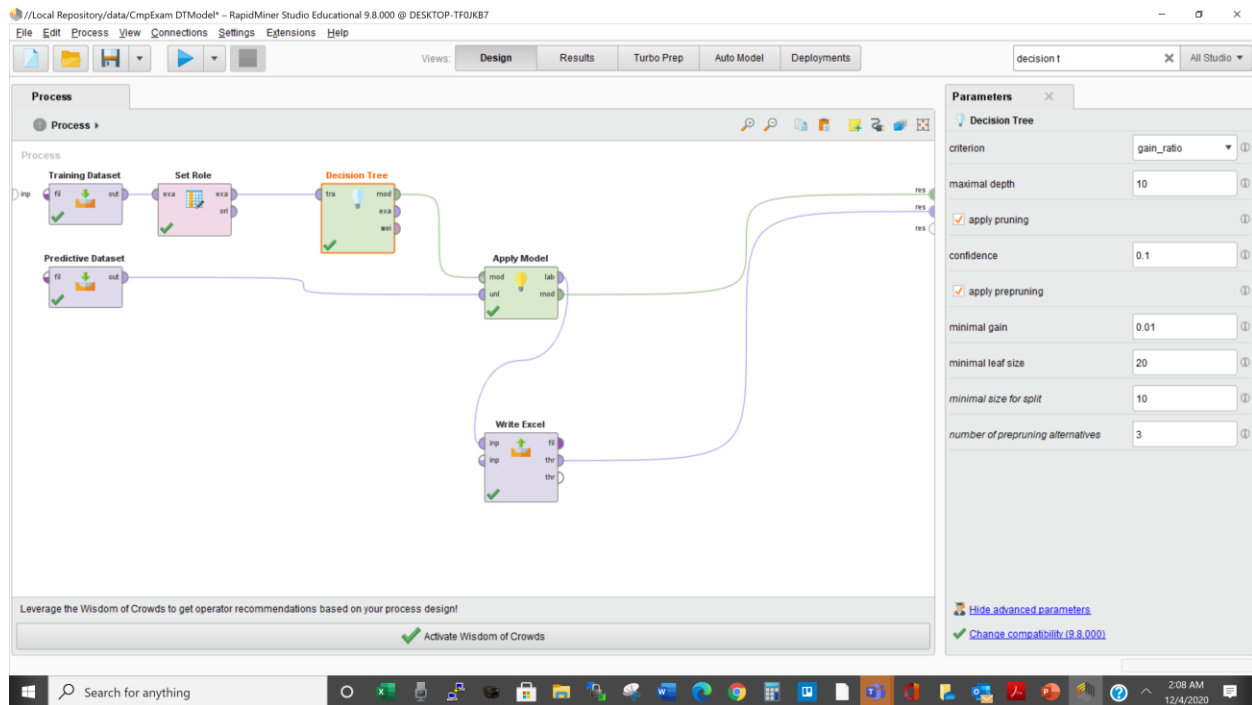
- Dawson, P., & Andriopoulos, C. (2017). *Managing Change, Creativity and Innovation* (3rd ed.). SAGE Publications Ltd.
- Decision Tree Analysis: Choosing by Projecting “Expected Outcomes.” (2020). Mind Tools. <https://www.mindtools.com/dectree.html#:~:text=Decision%20trees%20provide%20a%20effective,the%20probabilities%20of%20achieving%20them>
- Kaplan, R.S. and Norton, D.P. (1992) The balanced scorecard – measures that drive performance. *Harvard Business Review* 70 (1): 71–9.
- Landis, T. (2020). Customer Retention Marketing vs. Customer Acquisition Marketing. OutboundEngine. <https://www.outboundengine.com/blog/customer-retention-marketing-vs-customer-acquisition-marketing/>
- Market Segmentation: Everything to Know in 2020* //. (2020, October 14). Qualtrics. <https://www.qualtrics.com/experience-management/brand/what-is-market-segmentation/>
- Predictive Analytics: What it is and why it matters. (2020). SAS. https://www.sas.com/en_us/insights/analytics/predictive-analytics.html#:~:text=Predictive%20analytics%20are%20used%20to,forecast%20inventory%20and%20manage%20resources.
- Press, G. (2016, March 23). Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. Retrieved from <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#722588936f63>
- Punt, D. (2020, October 1). *The world's oldest people and their secrets to a long life*. Guinness World Records. <https://www.guinnessworldrecords.com/news/2020/10/the-worlds-oldest-people-and-their-secrets-to-a-long-life-632895>
- Target Market (2020). Wikipedia. https://en.wikipedia.org/wiki/Target_market
- University of Eastern Finland. (n.d.). Machine learning. Retrieved from <http://www.cs.joensuu.fi/~whamalai/skc/ml.html>
- Wikipedia contributors. (2020, October 29). *Market segmentation*. Wikipedia. https://en.wikipedia.org/wiki/Market_segmentation

Appendices

Appendix 1.

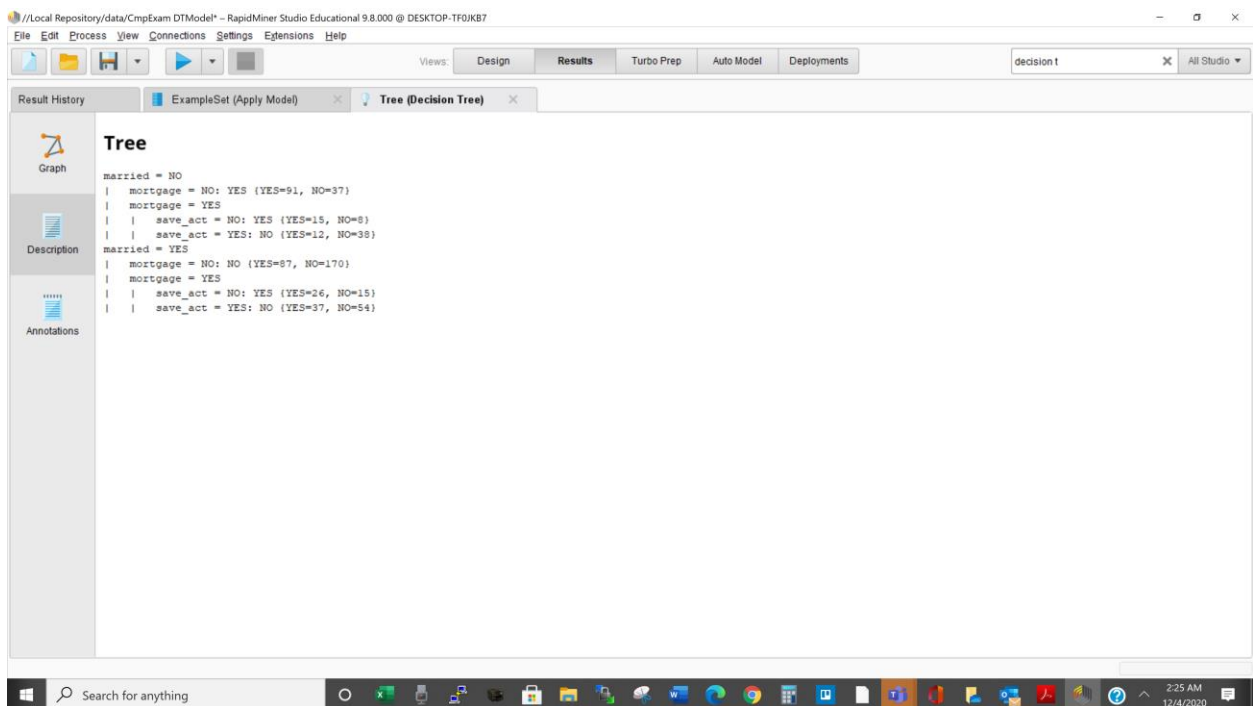
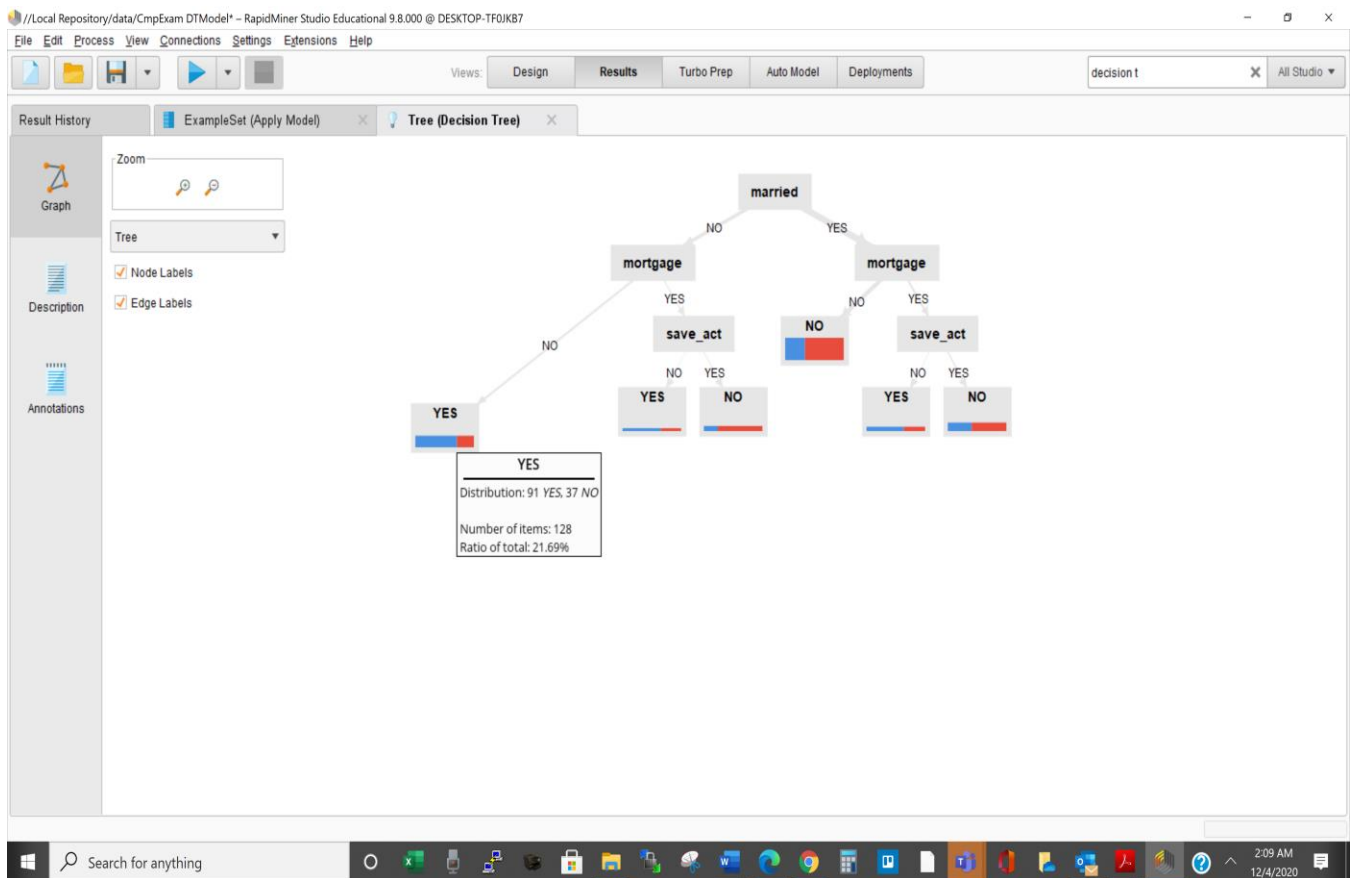
Using response as attribute name for a label

Decision Tree (a)



The image shows the RapidMiner Studio Results View. The **ExampleSet (Apply Model)** tab is selected, displaying the output of the **Decision Tree** process. The table shows 200 examples with 14 columns: Row No., prediction(...), confidence(...), confidence(...), id, age, sex, region, married, car, save_act, current_act, mortgage, and response.

Row No.	prediction(...)	confidence(...)	confidence(...)	id	age	sex	region	married	car	save_act	current_act	mortgage	response
1	NO	0.407	0.593	ID12701	23	MALE	INNER_CITY	YES	YES	YES	NO	YES	?
2	NO	0.240	0.760	ID12702	30	MALE	RURAL	NO	NO	YES	NO	YES	?
3	YES	0.711	0.289	ID12703	45	FEMALE	RURAL	NO	YES	YES	YES	NO	?
4	NO	0.407	0.593	ID12704	50	MALE	TOWN	YES	NO	YES	NO	YES	?
5	NO	0.339	0.661	ID12705	41	FEMALE	INNER_CITY	YES	YES	YES	YES	NO	?
6	NO	0.240	0.760	ID12706	20	MALE	INNER_CITY	NO	NO	YES	YES	YES	?
7	NO	0.407	0.593	ID12707	46	FEMALE	RURAL	YES	YES	YES	NO	YES	?
8	NO	0.407	0.593	ID12708	50	FEMALE	INNER_CITY	YES	YES	YES	NO	YES	?
9	YES	0.711	0.289	ID12709	42	MALE	INNER_CITY	NO	YES	YES	NO	NO	?
10	NO	0.339	0.661	ID12710	57	FEMALE	TOWN	YES	YES	NO	YES	NO	?
11	NO	0.407	0.593	ID12711	63	FEMALE	INNER_CITY	YES	NO	YES	NO	YES	?
12	YES	0.652	0.348	ID12712	26	FEMALE	INNER_CITY	NO	YES	NO	YES	YES	?
13	NO	0.339	0.661	ID12713	62	FEMALE	RURAL	YES	YES	YES	YES	NO	?
14	NO	0.339	0.661	ID12714	26	FEMALE	SUBURBAN	YES	YES	YES	YES	NO	?
15	NO	0.339	0.661	ID12715	19	MALE	RURAL	YES	NO	YES	YES	NO	?
16	YES	0.634	0.366	ID12716	44	MALE	TOWN	YES	NO	NO	NO	YES	?
17	NO	0.407	0.593	ID12717	32	FEMALE	INNER_CITY	YES	YES	YES	YES	YES	?
18	NO	0.239	0.661	ID12718	56	FEMALE	RURAL	YES	YES	YES	NO	NO	?



Decision Tree (b)


```

| | | | | save_act = NO
| | | | | | age > 22.500
| | | | | | | region = INNER_CITY
| | | | | | | | age > 54.500: YES {YES=6, NO=0}
| | | | | | | | age ≤ 54.500: NO {YES=8, NO=9}
| | | | | | | region = RURAL
| | | | | | | | age > 46: NO {YES=0, NO=2}
| | | | | | | | age ≤ 46: YES {YES=5, NO=0}
| | | | | | | region = TOWN
| | | | | | | | sex = FEMALE: NO {YES=2, NO=3}
| | | | | | | | sex = MALE: YES {YES=4, NO=1}
| | | | | | | age ≤ 22.500: YES {YES=2, NO=0}
| | | | | save_act = YES
| | | | | | age > 31.500: YES {YES=48, NO=12}
| | | | | | age ≤ 31.500
| | | | | | | current_act = NO: NO {YES=0, NO=3}
| | | | | | | current_act = YES: YES {YES=7, NO=4}
| | | mortgage = YES
| | | | age > 24.500
| | | | | save_act = NO
| | | | | | age > 26.500: YES {YES=14, NO=4}
| | | | | | age ≤ 26.500: NO {YES=1, NO=2}
| | | | | save_act = YES
| | | | | | age > 28
| | | | | | | age > 31.500: NO {YES=9, NO=23}
| | | | | | | age ≤ 31.500: YES {YES=2, NO=0}
| | | | | | | age ≤ 28: NO {YES=0, NO=3}
| | | | | age ≤ 24.500: NO {YES=0, NO=13}
| | married = YES
| | | mortgage = NO
| | | | age > 62.500
| | | | | sex = FEMALE
| | | | | | age > 64.500: NO {YES=0, NO=2}
| | | | | | age ≤ 64.500: YES {YES=3, NO=1}
| | | | | | sex = MALE: YES {YES=4, NO=0}
| | | | | age ≤ 62.500
| | | | | | age > 57.500: NO {YES=1, NO=11}
| | | | | | age ≤ 57.500
| | | | | | | age > 19.500
| | | | | | | | age > 56.500: YES {YES=1, NO=1}
| | | | | | | | age ≤ 56.500: NO {YES=70, NO=144}
| | | | | | | age ≤ 19.500
| | | | | | | | car = NO: YES {YES=1, NO=1}
| | | | | | | | car = YES: NO {YES=0, NO=4}
| | | mortgage = YES
| | | | age > 63.500
| | | | | car = NO: YES {YES=1, NO=1}
| | | | | car = YES: NO {YES=0, NO=3}
| | | | age ≤ 63.500
| | | | | save_act = NO
| | | | | | age > 58.500: YES {YES=3, NO=0}
| | | | | | age ≤ 58.500
| | | | | | | region = INNER_CITY: YES {YES=11, NO=5}
| | | | | | | region = RURAL: YES {YES=3, NO=1}
| | | | | | | region = SUBURBAN: YES {YES=4, NO=1}
| | | | | | | region = TOWN
| | | | | | | | sex = FEMALE: NO {YES=0, NO=4}
| | | | | | | | sex = MALE: YES {YES=3, NO=3}
| | | | | save_act = YES
| | | | | | age > 20.500
| | | | | | age > 37.500

```

```

| | | | | | | | | age > 39.500: NO {YES=24, NO=27}
| | | | | | | | | age ≤ 39.500: YES {YES=4, NO=0}
| | | | | | | | | age ≤ 37.500: NO {YES=8, NO=19}
| | | | | | | | | age ≤ 20.500: NO {YES=0, NO=2}
age ≤ 18.500: NO {YES=1, NO=9}

```

Logistic Progression

Local Repository\data\CmpExam DTModel* - RapidMiner Studio Educational 9.8.000 @ DESKTOP-TF0JKB7

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments

decision t X All Studio

Result History

ExampleSet (Apply Model) X Logistic Regression Model (Logistic Regression) X

Open in Turbo Prep Auto Model

Filter (200 / 200 examples): all

Row No.	prediction(r...	confidence(...	confidence(...	age	sex	region	married	car	save_act	current_act	mortgage	response
1	NO	0.257	0.743	23	MALE	INNER_CITY	YES	YES	YES	NO	YES	?
2	NO	0.493	0.507	30	MALE	RURAL	NO	NO	YES	NO	YES	?
3	YES	0.588	0.412	45	FEMALE	RURAL	NO	YES	YES	YES	NO	?
4	NO	0.351	0.649	50	MALE	TOWN	YES	NO	YES	NO	YES	?
5	NO	0.341	0.659	41	FEMALE	INNER_CITY	YES	YES	YES	YES	NO	?
6	NO	0.428	0.572	20	MALE	INNER_CITY	NO	NO	YES	YES	YES	?
7	NO	0.344	0.656	46	FEMALE	RURAL	YES	YES	YES	NO	YES	?
8	NO	0.342	0.658	50	FEMALE	INNER_CITY	YES	YES	YES	NO	YES	?
9	YES	0.584	0.416	42	MALE	INNER_CITY	NO	YES	YES	NO	NO	?
10	YES	0.504	0.496	57	FEMALE	TOWN	YES	YES	NO	YES	NO	?
11	NO	0.402	0.598	63	FEMALE	INNER_CITY	YES	NO	YES	NO	YES	?
12	YES	0.526	0.474	26	FEMALE	INNER_CITY	NO	YES	NO	YES	YES	?
13	NO	0.489	0.511	62	FEMALE	RURAL	YES	YES	YES	YES	NO	?
14	NO	0.349	0.651	26	FEMALE	SUBURBAN	YES	YES	YES	YES	NO	?
15	NO	0.290	0.710	19	MALE	RURAL	YES	NO	YES	YES	NO	?
16	NO	0.420	0.580	44	MALE	TOWN	YES	NO	NO	NO	YES	?
17	NO	0.268	0.732	32	FEMALE	INNER_CITY	YES	YES	YES	YES	YES	?
18	NO	0.431	0.569	56	FEMALE	RURAL	YES	YES	YES	NO	NO	?

ExampleSet (200 examples, 3 special attributes, 9 regular attributes)

Search for anything

1:59 PM 12/4/2020

Logistic Regression Model

Warning:

Removed collinear columns [region.INNER_CITY, region.SUBURBAN, region.RURAL, sex.MALE, married.NO, car.YES, save_act.NO, current_act.NO, mortgage.YES, age]

Model Metrics Type: BinomialGLM

Description: N/A

model id: rm-h2o-model-logistic_regression-5

frame id: rm-h2o-frame-logistic_regression-5

MSE: 6.159724E-16

RMSE: 2.4818792E-8

R^2: 1.0

AUC: 1.0

pr_auc: 1.0

logloss: 2.4818792E-8

mean_per_class_error: 0.0

default threshold: 1.0

CM: Confusion Matrix (Row labels: Actual class; Column labels: Predicted class):

	YES	NO	Error	Rate
YES	268	0	0.0000	0 / 268
NO	0	322	0.0000	0 / 322

Totals 268 322 0.0000 0 / 590

Gains/Lift Table (Avg response rate: 54.58 %, avg score: 100000.00 %):

Group	Cumulative Data	Fraction	Lower Threshold	Life	Cumulative Life	Response Rate	Score	Cumulative Response Rate	Cumulative Score	Capture Rate	Cumulative Capture Rate	Gain	Cumulative Gain
1	0.99930598	1000.000000	1.001698	1.001698	1.001698	0.546689	1000.000000	0.546689	1000.000000	1.000000	1.000000	0.149779	0.149779
2	0.00069402	1.00000000	1000.000000	0.000000	1.000000	0.000000	1000.000000	0.545763	1000.000000	0.000000	1.000000	-100.000000	-100.000000

null DOF: 589.0

residual DOF: 0.0

null deviance: 812.96436

residual deviance: 2.9286175E-5

GLM Model (summary):

Family	Link	Regularization	Number of Predictors	Total Number of Active Predictors	Number of Iterations	Training Frame
binomial	logit	None	599	589	50	rm-h2o-frame-logistic_regression-5

Scoring History:

timestamp	duration	iterations	negative_log_likelihood	objective
2020-12-04 13:30:05	0.000 sec	0	406.48219	0.68895
2020-12-04 13:30:05	0.048 sec	1	73.71467	0.12494
2020-12-04 13:30:05	0.049 sec	2	24.75892	0.04196
2020-12-04 13:30:05	0.050 sec	3	8.84524	0.01499

2020-12-04 13:30:05	0.050 sec	4	3.22054	0.00546
2020-12-04 13:30:05	0.051 sec	5	1.18034	0.00200
2020-12-04 13:30:05	0.051 sec	6	0.43363	0.00073
2020-12-04 13:30:05	0.052 sec	7	0.15944	0.00027
2020-12-04 13:30:05	0.052 sec	8	0.05865	0.00010
2020-12-04 13:30:05	0.053 sec	9	0.02157	0.00004

2020-12-04 13:30:05	0.079 sec	41	0.00002	0.00000
2020-12-04 13:30:05	0.080 sec	42	0.00002	0.00000
2020-12-04 13:30:05	0.080 sec	43	0.00002	0.00000
2020-12-04 13:30:05	0.081 sec	44	0.00002	0.00000
2020-12-04 13:30:05	0.081 sec	45	0.00002	0.00000
2020-12-04 13:30:05	0.082 sec	46	0.00002	0.00000
2020-12-04 13:30:05	0.083 sec	47	0.00002	0.00000
2020-12-04 13:30:05	0.083 sec	48	0.00002	0.00000
2020-12-04 13:30:05	0.084 sec	49	0.00002	0.00000
2020-12-04 13:30:05	0.084 sec	50	0.00001	0.00000

H2O version: 3.30.0.1-rm9.7.1

