**PROJECT REPORT**

**USING ADVANCED DATA ANALYSIS METHODS TO BOOST**

**NON-PROFIT ORGANIZATION EFFICIENCY**

**Introduction:**

In the world of non-profit organizations, direct advertisements are extremely effective at raising donations while involving previous donors. However, to make these campaigns successful, we need an analysis of previous donors and their behavior.

To work on this project, we are working with a client to develop the classification models to identify the donors accurately and regression models to predict the expected amount of donation. The main goal is to maximize the expected net profit by analyzing the recipients who will donate while minimizing the costs of non-donors.

**Business Understanding:**

The main issue of this is improper use of resources which results in net loss. This project uses predictive modeling techniques for organizations to identify the target individuals who can donate and estimate their donation amounts. The primary goal of this organization is to change its approach in identifying potential donors. The current process has produced just over a 10% response rate, with an average donation of $14.50 barely covering the $2.00 cost per letter. Understanding the characteristics of the donors can help to determine the profiles. This vital step is essential for raising each campaign's net profit and in changing a financial hardship into a profitable activity.

**Analytical Questions:**

1. How does past donation behavior relate to the likelihood and amount of future donations?
2. Analyze existing data and build a classification model to identify the potential donor so that the expected net profit is maximized.
3. Analyses data to predict the expected gift amount from the donor.

**Data Understanding/Exploratory Data Analysis:**

**Data Understanding:**

In our dataset, we are having 6002 rows, and each column has different data types.

* ID which is a unique one - alphanumeric type
* Region - Categorical
* Ownd {Ownership status} - Binary
* Kids - Numeric
* Inc {income} - Numeric
* Sex - Categorical
* Wlth {Wealth} - Numeric
* Hv{Home value} - Numeric
* Incmed {Median Income} - Numeric
* Incavg {Average Income} - Numeric
* Low - Binary
* Npro {Non-profit level} - Numeric
* Gifdol {Amount(Dollar) of Gifts donated} - Numeric
* Gifl {Largest gift amount} - Numeric
* Gifr {Recent gift amount} - Numeric
* Mdon {Median donation amount} - Numeric
* Lag - Numeric
* Donr {Donor} - Binary
* Damt - Numeric

If we look into the target variables among all, we can consider “Donr” and “Damt” variables. ‘donr’ is the target variable for the classification model by this we can estimate the donation amounts and next ‘damt’ is the target variable for the regression model by this we can estimate the donor categorization.

**Exploratory Data Analysis (EDA):**

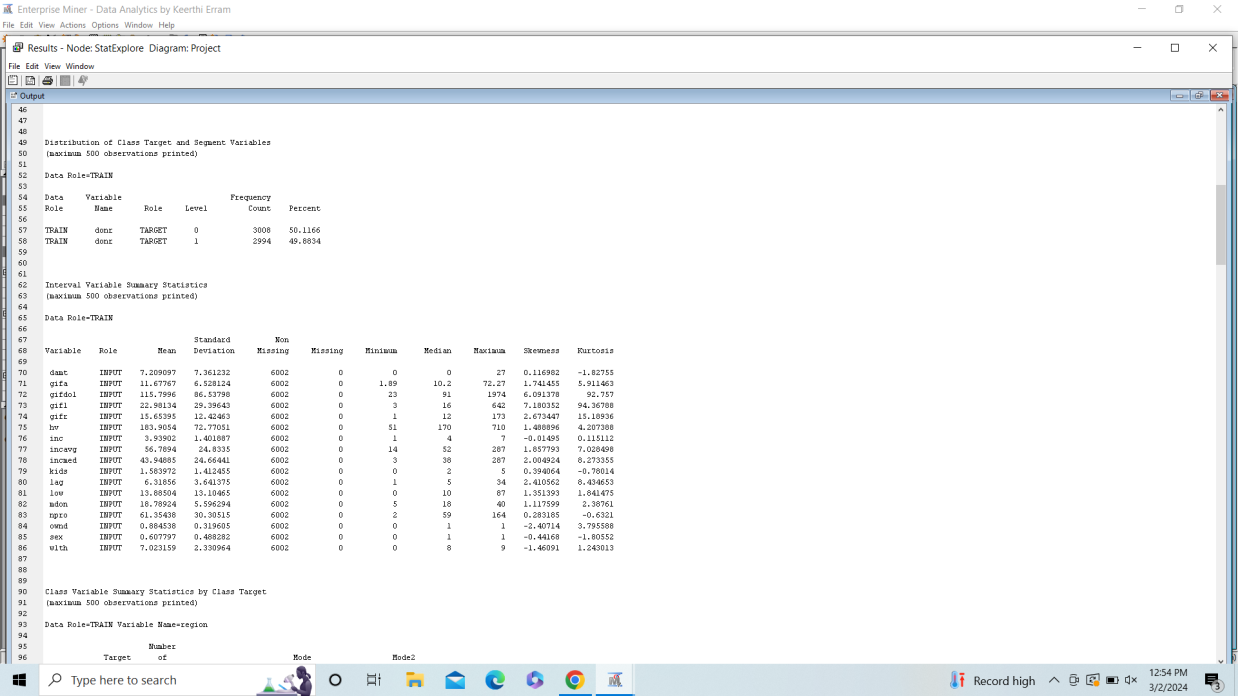
We know EDA is the main step before we move onto the data analysis project. Initially, we have considered “StatExplore” and “Graph Explore” nodes for the analysis of EDA.

We have used “StatExplore” node because it involves the statistical analysis of the data and moreover it focuses on summary statistics and other numerical measures and “Graph Explore” node is used to visually explore the data and identify patterns for further analysis and for decision-making.

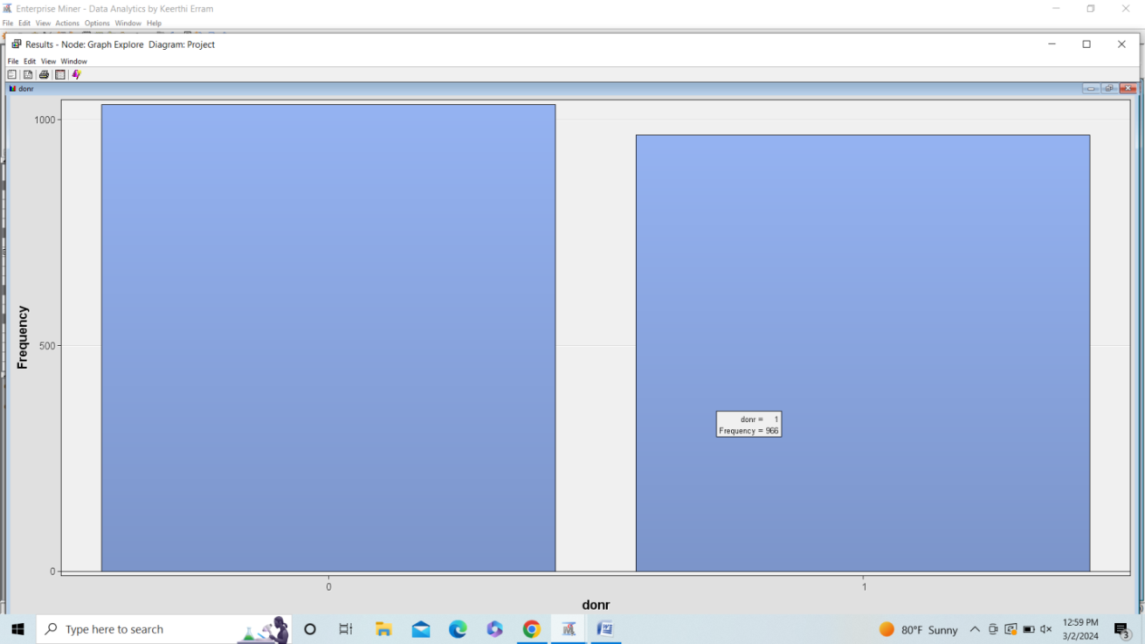
If we look into the StatExplore node, considering “donr” variable as target, we get the result as below image. In the variable summary section, we can see that the dataset contains one nominal input variable, one ID variable {i.e., Id variable}, 17 interval input variables {i.e., Incavg, Npro, Gifl Gifdr, etc…}, and one nominal target variable {i.e., donr variable}. In the distribution of class target and segment variables, we can see that the count and percentages of 0s and 1s in the “donr” variable.

When it comes to interval variable summary statistics section, it shows the statistics like mean, standard deviation, minimum, maximum, missing, skewness, and kurtosis for each variable. If we look into the missing column, where we have seen that there are no missing values in the dataset.

For example, consider "hv" among all the variables and observe the mean, standard deviation, minimum, maximum, skewness, and kurtosis. We have 1.488896 skewness, which shows positive skewness and a right-skewed distribution. Overall, these statistics indicate that the "hv" variables have a right-skewed distribution. Similarly, when we consider the "ownd" variable, these statistics indicate a left-skewed distribution, similarly with the "wlth" variable. We can look into all the summary statistics for all the variables and here we considered only 3 variables as an example.



If we look into the “Graph Explore” node results, we can see that the target variable has 2 levels one is “0” and the other one is “1”. The below graph shows the frequency of donr variable. For donr with 0 level has the frequency count of 1034 and for donr with 1 level has the frequency count of 966. Overall, the target variable “donr” is generally uniform distribution and this suggests more analysis can be done to better understand its interaction with other variables and make predictions from the data.



**Data preparation:**

Data preparation is the act of cleaning and transforming raw data prior to processing and analysis. This is one of the significant steps prior to processing and often involves reformatting data, making corrections to data, and combining datasets to enrich data. Data preparation also includes data partition and data transformation steps.

**Data upload:**

Drag the “File Import” node and import the file “nonprofit” in the Diagram. Then edit the variable property for “Classification Algorithms” initially.

A screenshot of a computer

Description automatically generated

**Data Explore:**

StatExplore node– This node used

1. To see potential relationships between all the variables.
2. To see if any missing points are in the data and to see if the data is clean.

Made a note that Variable Worth and the Chi-Square plot and this shows the **“kids**” variable has the potential largest relationship with target variable donr.

A screenshot of a computer

Description automatically generated

Output window shows that **there is no missing value in the data.** This confirms that data is clean. Data cleaning or wrangling activity is not required on the data.

A screenshot of a computer

Description automatically generated

**Data Partition:**

The purpose of database partitioning is to improve the database's performance, scalability, and availability. By using “Partition node” we can partition the data and kept “Training portion 40%, Validation portion 30%, and Test portion 30%”.

A screenshot of a computer

Description automatically generated

In the “Exported Data” property we have verified that train, validation and test data is successfully partitioned.

A screenshot of a computer

Description automatically generated

**Transform Variables:**

In KNN algorithm, it is crucial to standardize variables to put all the internal variables on the same scale. If interval data are on different scales, then it causes problems when the algorithm goes to calculate the distances by imposing a bias from the different variables. This removes the issue of the influence of the scale when determining these distances.

Next, we have dragged “Transform Variables” node and set method to “Range Standardization” for all interval variables.

A screenshot of a computer

Description automatically generated

After running the transform variable node, all internal variables will have a minimum of 0 and a maximum of 1. Below results screen, shows “Formula” column and also the transformation that was done to standardize the variable. The Missing column shows that there is no missing value in the data.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Methods / Modeling – Classification Algorithm:**

Developed a classification model for the “DONR” variable using ‘nonprofit’ data. We decided to use the algorithm below, because these algorithms are considered as better models for supervised data.

1. **Decision Trees & Random Forest –**Big Idea behind selecting decision tree model is Decision trees repeatedly split the records into parts to achieve maximum homogeneity within the new parts in terms of the target values. Also, it does not require standardization or normalization. Random forest selected because it adds extra randomness. It searches only in a random subset, rather than the entire set, of the predictors for growing the tree at each split. Typically, Classification and prediction with Random Forest is more accurate.
2. **Logistic Regression –** Logistic Regression model selected for classification algorithm because it is one of the popular and effective models for predicting binary variables. Logistic regression does not require normalization/standardization.
3. **KNN –** Big idea behind selecting KNN model is that it is one of the effective Non-parametric predictive modeling techniques for categorical or continuous targets. KNN model is Data-driven, not model-driven because of it there is no pre assumption on the data. It works well on classification and regression model.

**Classification Algorithm 1 – Decision Tree:**

We have created 2 decision trees and Random Forest model to predict the data. Those are:

1. 6 Branch 20 Deep
2. 3 Branch 10 Deep
3. HP Forest

Then we connected all the 3 nodes to “Control Point” node. This control point node allows for all paths leading to that node to be run at the same time. Also control point can also simplify and reduce the connections between process flow steps that have multiple interconnected nodes.

**A screenshot of a computer

Description automatically generated**

**Result - Model Comparison:**

Then we have used the “Model Comparison” node and set the selection statistic to “Misclassification Rate”. Next change selection table to “Test” as we need to do the comparison and that has to be done on Test data.

Below Fit Statistics show that Random forest is the preferred model as it has the **lowest misclassification rate on test data which is 0.10427.**

**A close-up of a number

Description automatically generated**

**Classification Algorithm 2 – Logistic Regression:**

By using, multiple logistic regression stepwise approach and select the predictor variables that needs to enter and stay in the model if they are at the set significant level.

Change the Logistic Regression property as mentioned below:

* Change Selection Model to “Stepwise”.
* Change Use Selection Defaults to “No” as we set Significance Level to 0.10.
* Change the Maximum Number of Steps to “20”.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Run the Logistic Regression node and made a note of results.

A screenshot of a computer

Description automatically generated

Fit Statistics table shows that if the Target variable donr is selected and if the test data is noticed then Misclassification Rate is coming as 0.138103. This is the closest value from the decision tree algorithms.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Train | Validation | Test |
| Misclassification Rate | 0.157083 | 0.17065 | 0.138103 |

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

**Classification Algorithm 3 – KNN:**

As mentioned earlier, normalization and standardization required on the data for KNN model before running the algorithm. Therefore, we will connect MBR node to transform variable. In Transform variable already all the internal variables are range standardized.

Dragged 10 MBR node from Model menu to check the best nearest distance model. Number of neighbors set to 2 for MBR =1 and then it raised by 1 for each node. MBR=10 has 11 as number of neighbors value. Then Run model comparison.

A diagram of a model comparison

Description automatically generated

Based on the output provided by the optimal k value that is 9. The misclassification rate is coming as 0.19190 on the node MBR=9 on the Test data which is better than any other MBR nodes.

A screenshot of a computer screen

Description automatically generated

**Evaluation for Classification Algorithms:**

As we have the result of all the models then let’s evaluate model using Model Comparison node. Performing “Model Comparison” on Test data, with all the classification algorithms decision tree, logistic regression and KNN.

**A screenshot of a computer

Description automatically generated**

Below fit statistics show that Random Forest is preferred model for classification because it has lowest misclassification rate = 0.10427 on Test data. Average Squared error comes as 0.09582 which is better than any other model.

A screenshot of a number of numbers

Description automatically generated

**Revisiting Business Objective**

With implementing various classification algorithms, we came up with a robust classification algorithm which is a Decision Tree-Random Forest. This model will predict the donors 91% accuracy. It is not cost-effective to mail everyone because the expected profit from each mailing is 14.50 x 10% – 2 = -$0.55. Therefore, we trained the model with trained data, validated it validation data, and tested the results on Test data. There is no high variance between Train, Validation, and Test data so there is no overfitting issue. Our preferred model will correctly run on score data and provide the results. This model can effectively capture likely donors so that the expected net profit will be maximized.

**Deployment and Score New Data for Classification Algorithms:**

Model evaluation shows that Random Forest is the better model to predict the donr variables on score data. Added Score node in the diagram and imported “nonprofit\_score” excel to classify donr variable with classification model.

A screenshot of a computer

Description automatically generated

Made a note of “Score Data” results. Noted that “Random Forest” model considered evaluating the Score data. This model was considered as the preferred model to predict the data.

A screenshot of a computer

Description automatically generated

Summary Statistics of Output window shows that our best model (Random Forest) predicted 446 donors. There are 1561 non-donors are identified. These 446 predicted donors should receive a flyer so that the expected net profit is maximized.

A close-up of a data

Description automatically generated

Let’s export score data in excel for further analysis.

A screenshot of a computer

Description automatically generated

Explore the test data and confirm that prediction column is added by the model. Then right clicked on the data and extracted the Score prediction in excel.

A screenshot of a computer

Description automatically generated

**Business Insight and Applying the Findings**

**Identified Donors & Mailing Cost**

Decision Tree- Random forest which is our best model predicted that 446 predicted donors should receive a flyer. As per the model remaining 1561 are less likely to donate.  The mailing costs $2.00 to produce and send. $ 892 (446 donors \* $2 cost )will be the mailing cost to send emails to these identified donors.

**Limit mailing expenses and Maximized Profit**

Non-profit organization will save $3122 (1561\*2) by not sending emails to predict non donors. Also, the average donation is $14.50 so organization can expect $6467 (446\*14.50 avg donation) donations from these 446 identified donors.

Cost and profit calculated with Random forest

Potential Cost $4014= Potential donors 2007 \* $2 mailing cost

Identified cost $892 = Identified donors 446 \* $2 mailing cost

Cost Saving $3122 = Potential Cost $4014 - $892 Identified cost

Identified profit $6467 = Identified donors 446 \* 14.50 avg donation

**Methods / Modeling – Regression Algorithm:**

Developed a Regression model for the “DAMT” variable using ‘nonprofit’ data. We decided to use the algorithm below as these algorithms are considered as better models for supervised data.

1. **Linear Regression –** This is the most popular model to make predictions. The model is used to fit relationships between a numerical outcome variable (target variable) and set of predictors (input variables)
2. **Decision Trees & Random Forest –** Decision Tree goal is to predict an outcome using rules on a set of predictors. The output is set by the rules and rules are represented by the tree diagram. The Decision Tree helps interpret the associations in the data.
3. **KNN -** KNN is an algorithm that can be used for either classification or prediction models. KNN looks at the k nearest labeled point to take a majority vote for what class we should assign to X.

**Data upload:**

Drag the “File Import” node and import the file “nonprofit” in the Diagram. Then edit the variable property for “Regression Algorithms” initially by considering ‘damt’ variable as target variable.

A screenshot of a computer

Description automatically generated

**Regression Algorithm 1 – Linear Regression:**

We have used multiple linear regression stepwise approach by selecting predictor variables that enter and stay in the model if they are at the set significant level.

Change the Linear Regression property as per below.

* Change the Selection Model to “Stepwise”.
* Change Use Selection Defaults to “No” as we set the Significance Level to 0.10.
* Change the Maximum Number of Steps to “20”.

A screenshot of a computer

Description automatically generated

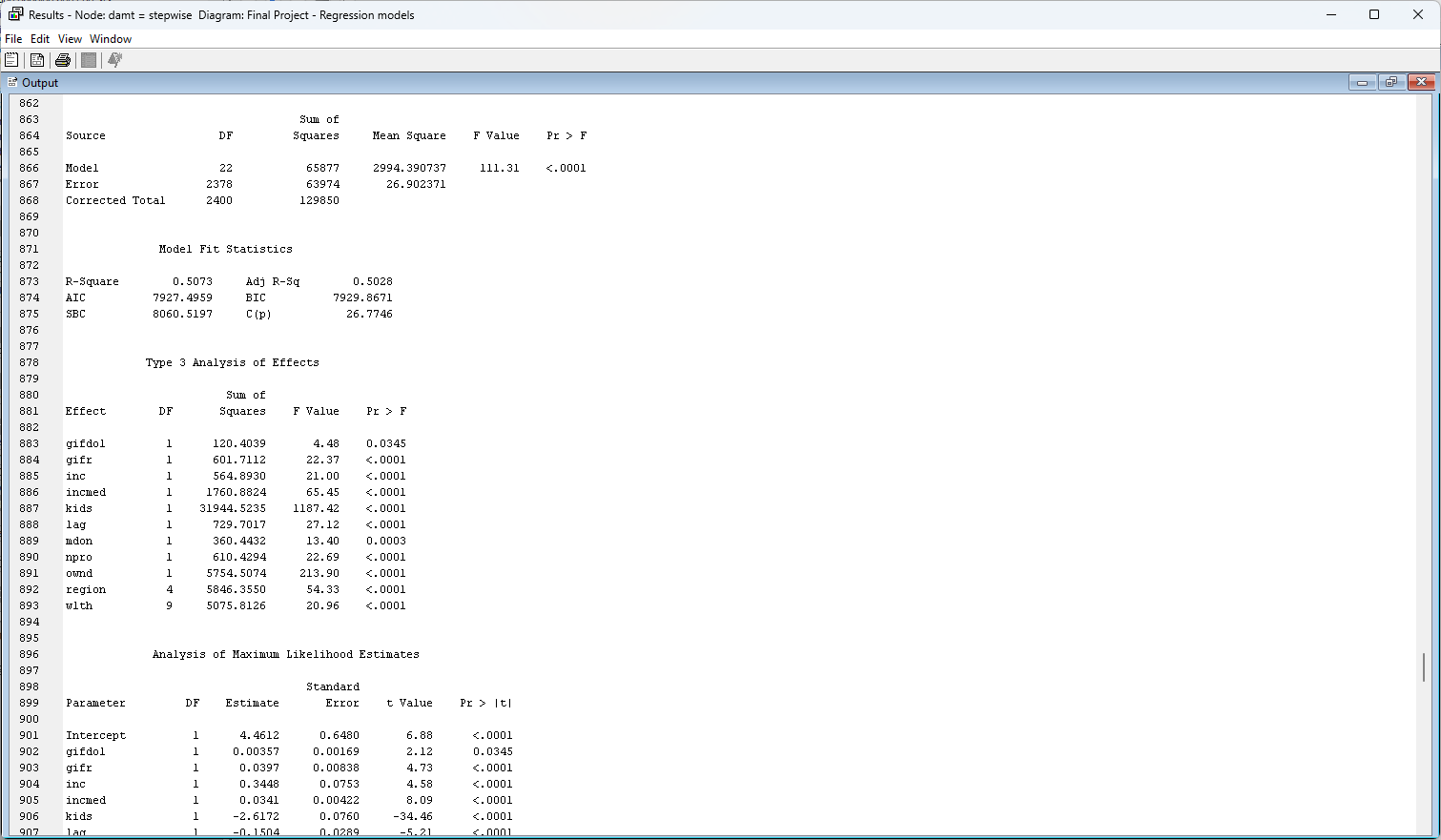
Run the Linear Regression node and made a note of results.

A screenshot of a computer

Description automatically generated

Fit Statistics table shows that if the target variable ‘damt’ is selected and if the test data is noticed then we have

* Average Squared Error is 26.51
* Validation Average Squared Error is 27.94
* Train Average Squared Error is 26.64



**Regression Algorithm 2 – Decision Tree:**

Created 2 decision Tree and Random Forest model to predict the data.

1. 6 Branch 20 Deep
2. 3 Branch 10 Deep
3. HP Forest

Then we have connected all 3 nodes to “Control Point” node. This Control Point node allows for all the paths leading to that node to be run at the same time. Also control point can also simplify and reduce the connections between process flow steps that have multiple interconnected nodes.

A screenshot of a computer

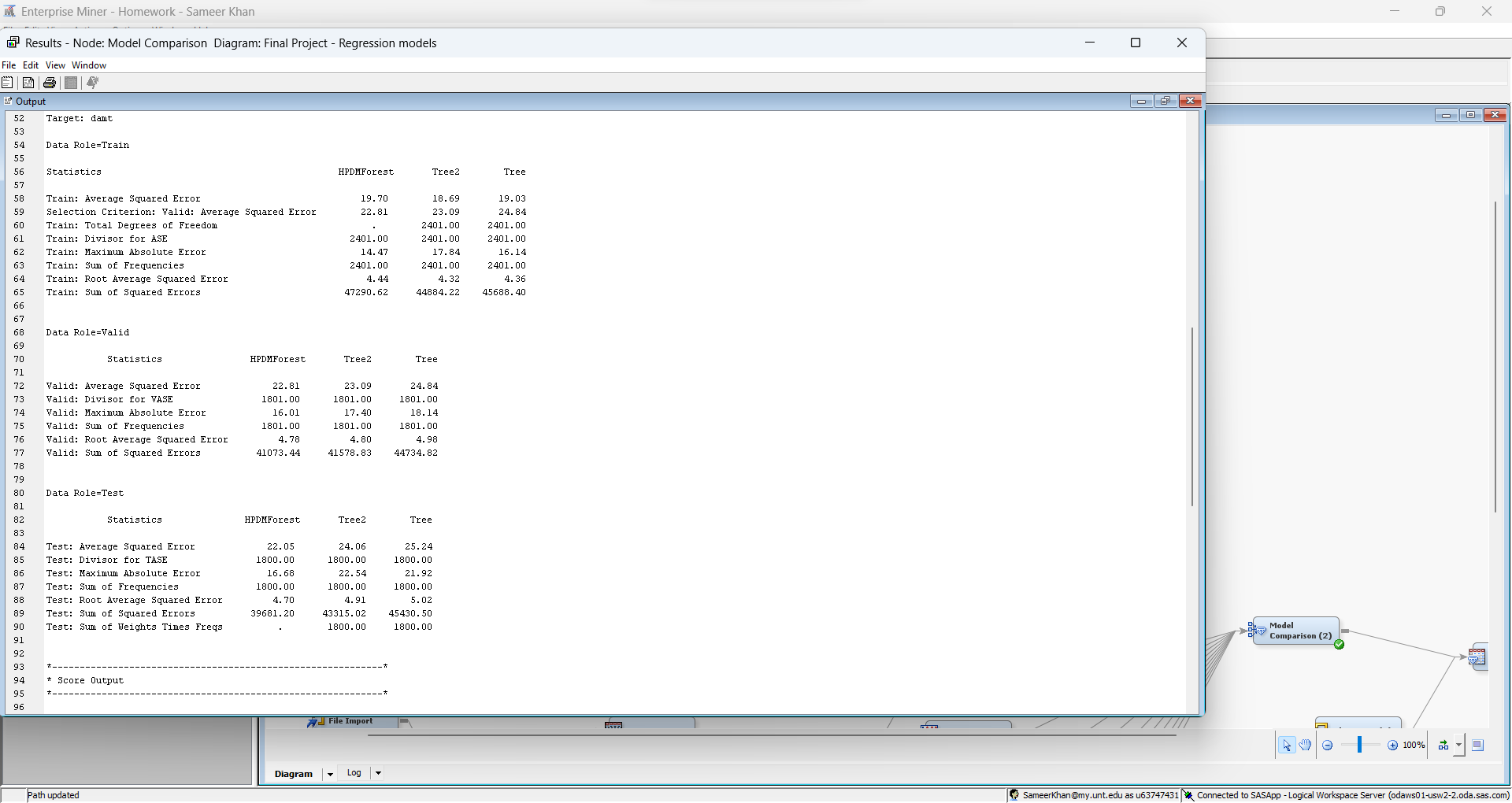
Description automatically generated

**Result - Model Comparison:**

Added Model Comparison node and set the Selection Statistic to “Average Squared Error”. Then change Selection Table to Test as we need comparison needs to be done on Test data.

Below Fit Statistics show that Random Forest is the preferred model as it has the **lowest Average Squared Error on Validation data which is 22.8059.** Next, if we look into the Train Average Squared Error: 19.6962.

A screenshot of a computer

Description automatically generated

**Regression Algorithm 3 – KNN:**

As mentioned earlier, normalization and standardization are required on the data for KNN model before running the algorithm. Therefore, we will connect MBR node to transform variable. In Transform variable already all the internal variables are range standardized.

Dragged 10 MBR node from Model menu to check the best nearest distance model. The number of neighbors is set to 2 for MBR =1 and then it raised by 1 for each node. MBR=10 has 11 as number of neighbors value. Then Run model comparison.

A screenshot of a computer

Description automatically generated

Based on the output provided, the optimal k value consists of 9. The Average Squared Error is coming as 32.8495 on the node MBR=9 on the Validation data which is better than any other MBR nodes and we have Train Average Squared Error: 24.9234.

A screenshot of a computer

Description automatically generated

**Evaluation for Regression Models:**

As we have the result of all the models then let’s evaluate the model using Model Comparison node. We have performed “Model Comparison” on Test data with all the Average Squared Error algorithms decision tree, linear regression and KNN.

A screenshot of a computer

Description automatically generated

Below fit statistics show that Random Forest is preferred model for regression because it has lowest Average Squared Error = 22.0451 on Test data.

A screenshot of a computer

Description automatically generated

**Revisiting Business Objective**

With implementing various Regression algorithms, we came up with a robust classification algorithm which is a Decision Tree-Random Forest. This model will predict the damt variable with lowest Average Squared Error = 22.0451. It is not cost-effective to mail everyone because the expected profit from each mailing is 14.50 x 10% – 2 = -$0.55. Therefore, we trained the model with trained data, validated it validation data, and tested the results on Test data. There is no high variance between Train, Validation, and Test data so there is no overfitting issue.

**Deployment and Score New Data for Regression Models:**

Model evaluation shows that Random Forest is the better model to predict the “damt” variables on score data. Added Score node in the diagram and imported “nonprofit\_score” excel to classify ‘damt’ variable with regression model.

A screenshot of a computer

Description automatically generated

Made a note of the Score Data results. Noted that Random Forest model considered evaluating the Score data. This model was considered as preferred model to predict the data.

A screenshot of a computer

Description automatically generated**A screenshot of a computer

Description automatically generated**

Let’s export score data in Excel for further analysis.

Explore the test data and confirmed that prediction column is added by the model. Then right clicked on the data and extracted the Score prediction in excel.

A screenshot of a computer

Description automatically generated

**Business Insight and Applying the Findings**

**Profit calculation**

The calculation of the Final Profit is determined by using the 446 donors from the classification model and multiplying by the mailing cost of $ 2.00. Giving us the total mailing expense cost of $ 892.

Identified cost $892 = Identified donors 446 \* $2 mailing cost

Then used a regression model to get the predicted damt of 446 donors. The below table shows that 446 donors will donate $4445.58 amount.

|  |  |  |
| --- | --- | --- |
| **Row Labels** | **Count** | **Predicted damt by regression model** |
| 0 – Non Donor | 1561 | $4902.97 |
| 1 - Donors | 446 | $4445.58 |
| **Grand Total** | **2007** | **$9348.55** |

Calculated the total cost and subtracted it by the sum of prediction of damt for 446 donors which was $ 4445.58 and getting the Final Profit of $ 3553.58.

**Final Profit (3553.58) = Predicted damt by regression model (4445.58) – (892) Mailing cost**

**Conclusion:**

As we know the primary goal of our project is to improve the net profit while minimizing the costs of non-donors. Among logistic regression, KNN, decision tree and random forest models that we performed for both classification and regression models, and we observed that the decision tree & random forest is the best suitable method to identify potential donors and predict the donation amounts that can maximize net profits.

**Key findings:**

* In the classification model we can observe that,

1. Logistic regression- The misclassification rate is 0.13810
2. Decision Tree-The misclassification rate is 0.10427
3. KNN-The misclassification rate is 0.19190

* Based on the above misclassification rates we came to conclusion that Decision tree is the best model. And also, we can observe that the potential donor’s frequency is 446 and the non-donors frequency is 1561.
* In the Prediction model we can observe that,

1. Linear Regression- The average squared error is 27.94
2. Decision Tree-The average squared error is 22.805
3. KNN- The average squared error is 32.849

* Based on the average squared error rates we came to conclusion that Decision tree is the best model. And also, we can observe that the potential donors’ frequency is 4445.58 and the non-donors frequency is 4902.98

**Limitations addressed for future research:**

* The donor behavior can be known accurately if we add more data regarding this to the existing one.
* Other than the classification and prediction models used in our analysis we can employ other models to find the efficiency of the data. This can also help to identify the best model with best accuracy.
* The behavior of donors might change over time which leads to the need for continuous changes in the data. This can also affect the predictions.

**Results for the Classification model**



**Results for the Regression model**



**References:**

* *References – Decision Tree Lecture, Logistic Regression Lecture, KNN Lecture*
* *References – Multiple Linear Regression Lecture, KNN Lecture, and Decision Tree Lecture.*
* *Author -Talend A QlikCompany (2022)* [*https://www.talend.com/resources/what-is-data-preparation/*](https://www.talend.com/resources/what-is-data-preparation/)