**DASARI HAREESH**

BanK Loan Analysis

# Introduction:

**Team :**

**Description:**

I made the decision to focus on the features that allow bank loans as well as the projection of bank loans based on the data set. Creating a prediction model and refining the attributes that are used to assign a label are the objectives of this activity.

**Problem Statement:** To specify the characteristics and create a model that aids in predicting bank loans

# Dataset Characteristics:

This dataset downloaded from <https://www.kaggle.com/>

The dataset has 1493 rows and 15 columns of data. There are multiple columns designated as model attributes, as well as one label column. The objective is to develop a strong model by training an algorithm on this dataset.

Feature Summary:

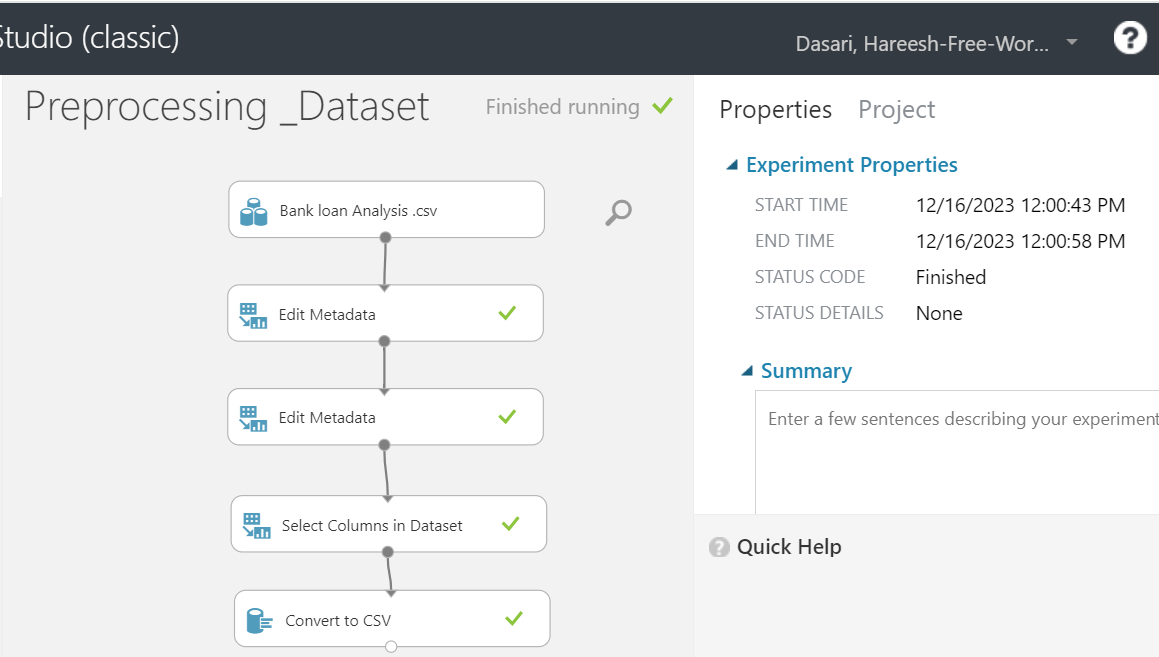
The characteristics are numerous. One Label Column (01 Label Column) and thirteen Feature Columns

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Feature/ Label** | **Count** | **Unique Value Count** | **Missing Value Count** | **Min** | **Max** | **Mean** |
| ID | 1492 | 1492 | 0 | 1 | 1492 | 752 |
| Age | 1492 | 44 | 0 | 21 | 67 | 46.29851 |
| Experience | 1492 | 41 | 0 | 0 | 42 | 22.02462 |
| Income | 1492 | 152 | 1 | 10 | 205 | 73.34083 |
| ZIP Code | 1492 | 387 | 0 | 90518 | 96651 | 99074.87 |
| Family | 1492 | 4 | 3 | 1 | 4 | 3.623616 |
| CCAvg | 1492 | 91 | 0 | 0 | 10 | 1.976654 |
| Education | 1492 | 4 | 2 | 1 | 4 | 1.862575 |
| Mortgage | 1492 | 225 | 0 | 1 | 720 | 53.87724 |
| Personal Loan | 1492 | 2 | 0 | 0 | 2 | 0.300734 |
| Securities Account | 1492 | 2 | 1 | 1 | 1 | 0.155738 |
| CD Account | 1492 | 2 | 0 | 0 | 1 | 0.08064 |
| Online | 1492 | 3 | 0 | 0 | 2 | 0.60547 |
| CreditCard | 1492 | 3 | 0 | 0 | 1 | 0.489556 |
| Loan Success (Label) | 1492 | 3 | 1 | 0 | 1 | 0.583326 |

# Steps Followed from Data Transformation:

**5 a) Pre-Processing Data**: This module comprises the data transformation section after data import. This module's data types can be changed, the values can be designated as category values, and any unnecessary columns can be removed.

* Edit Metadata – Float should be used as the numerical datatype.
* Edit Metadata – Designating each column as a feature
* Select Columns – Removing the optional columns and choosing just the mandatory ones
* Convert to CSV – To get the final file as a CSV,



The screenshot above depicts all of the preprocessed stages that are carried out as part of Data Transformation.

**5 b)**I have choose 3 regression models to perform on the dataset, here there are 3 models that are used for the analysis.

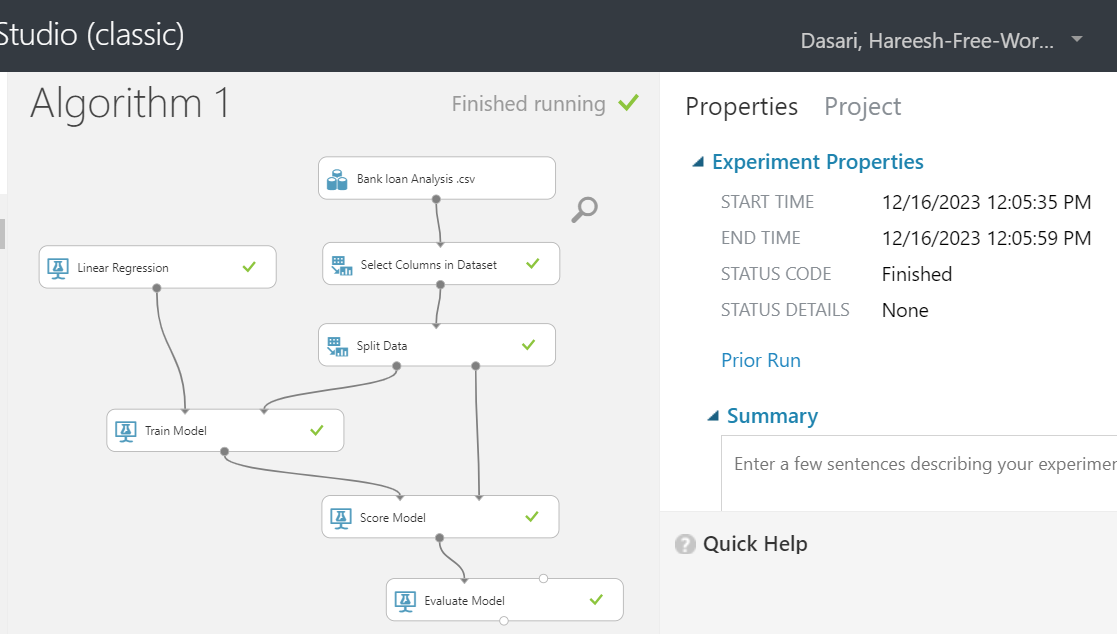
**1. Linear Regression Method:**

Ordinary Least Squares (OLS) is the method used, which is a statistical technique for fitting a linear regression model by minimising the sum of squared differences between observed and predicted values.

In addition, to prevent overfitting, L2 regularisation with a weight of 0.001 is used, the model includes an intercept term, uses a random seed number 6789 for reproducibility, and allows for unknown categorical values during the analysis.

Below are the default settings used of the linear regression

|  |  |
| --- | --- |
| Solution method | Ordinary least squares |
| L2 regularization weight | 0.001 |
| Include intercept term | Yes |
| Random seed number | 6789 |
| Allow unknown categorical values | Yes |



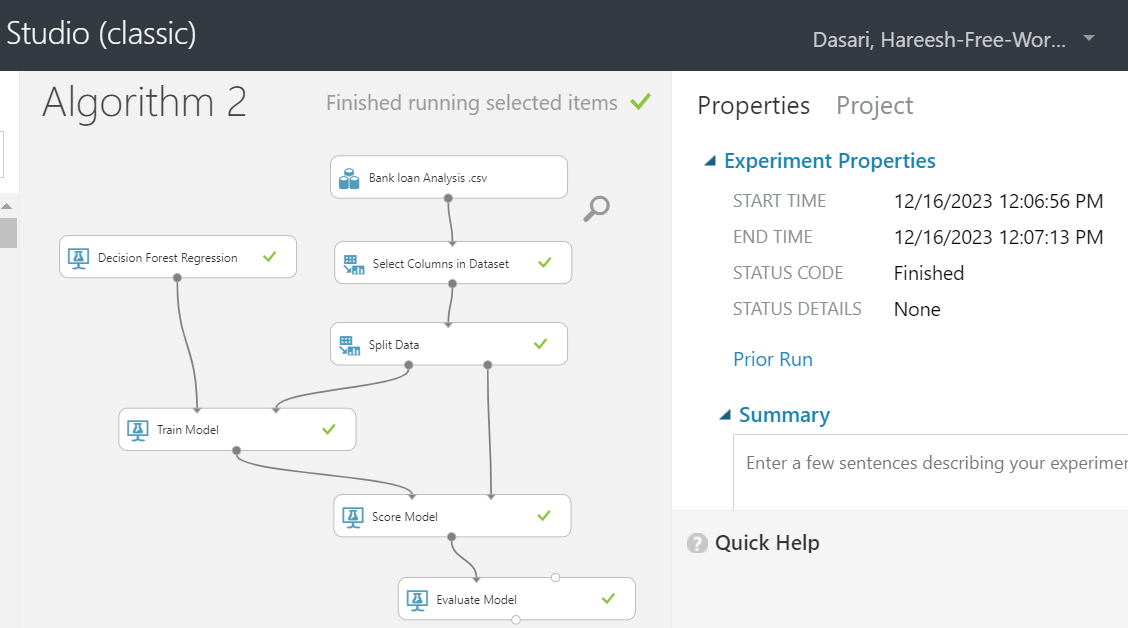
|  |  |
| --- | --- |
| **Metrics** | **Linear Regression** |
| Mean Absolute Error | 0.29627 |
| Root Mean Squared Error | 0.232197 |
| Relative Absolute Error | 0.492542 |
| Relative Squared Error | 0.532937 |
| Coefficient of Determination | **0.72937** |

2. **Decision Forest Regression:**

Default settings used for this model are below:

Bagging was selected as the resampling method, which entails training multiple models on different subsets of the dataset and combining their predictions to reduce overfitting and improve accuracy. This single-parameter trainer mode employs 8 decision trees. Each tree has a maximum depth of 32, 128 random splits per node, and a minimum of 1 sample per leaf node, with the goal of creating diverse trees and capturing complex relationships within the data while avoiding excessive depth or leaf node size restrictions.

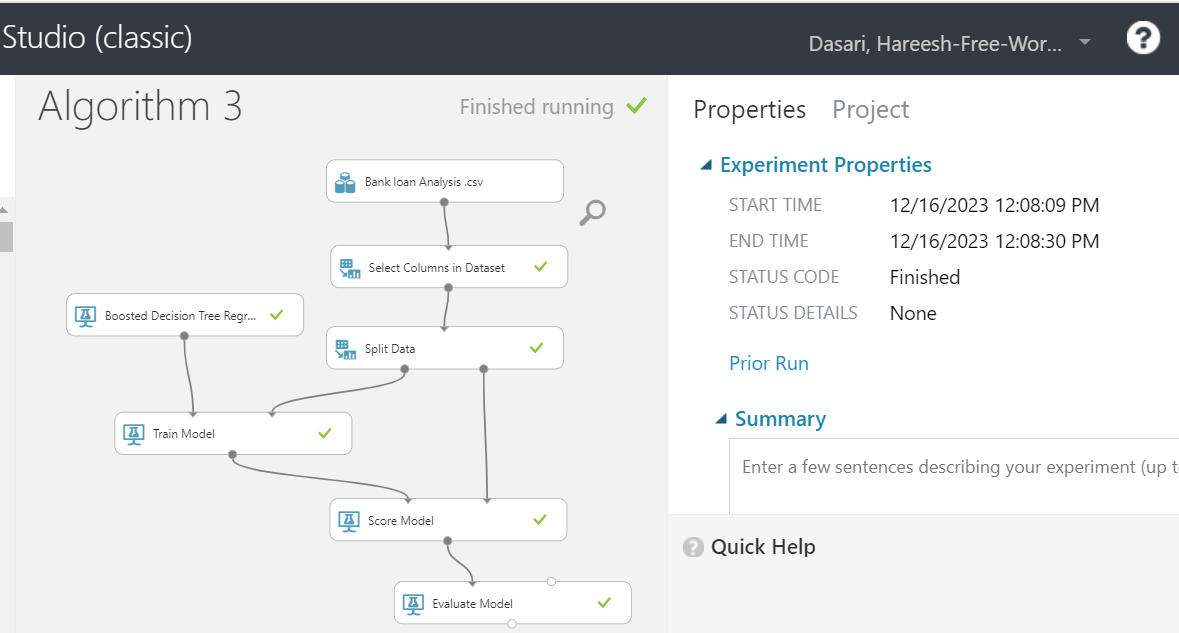
|  |  |
| --- | --- |
| Resampling method | Bagging |
| Create trainer mode | Single parameter |
| Number of decision trees | 8 |
| Maximum depth of the decision trees | 32 |
| Number of random splits per node | 128 |
| Minimum number of samples per leaf node | 1 |



|  |  |
| --- | --- |
| **Metrics** | **Decision Forest Regression** |
| Mean Absolute Error | 0.30344 |
| Root Mean Squared Error | 0.295451 |
| Relative Absolute Error | 0.207597 |
| Relative Squared Error | 0.62265 |
| Coefficient of Determination | 0.5065 |

3. **Boosted decision Tree Regression:**A gradient boosting model with a learning rate of 0.2 is used in the single-parameter trainer mode, constructing 200 trees with a maximum of 20 leaves per tree. Each leaf node requires a minimum of ten samples to ensure robustness against overfitting while allowing for unknown categorical levels and reproducibility via the specified random seed number 4567.

|  |  |
| --- | --- |
| Create trainer mode | Single parameter |
| Maximum number of leaves per tree | 20 |
| Minimum number of samples per leaf node | 10 |
| Learning rate | 0.2 |
| Number of trees constructed | 200 |
| Random seed number | 4567 |
| Allow unknown categorical levels | Yes |



|  |  |
| --- | --- |
| **Metrics** | **Boosted Decision tree regression** |
| Mean Absolute Error | 0.509444 |
| Root Mean Squared Error | 0.596065 |
| Relative Absolute Error | 1.019614 |
| Relative Squared Error | 1.422186 |
| Coefficient of Determination | 0.422186 |

**Conclusion:**

Linear Regression performs the best overall:

The Mean Absolute Error (0.29627) and Root Mean Squared Error (0.232197) are the lowest. It also has a reasonable balance of Relative Absolute Error (0.492542) and Relative Squared Error (0.532937), indicating moderate accuracy and explanatory power.

Furthermore, it has the highest Coefficient of Determination (R2) of 0.72937, indicating that it best explains the variance in the data among the three models. Given these metrics, the Linear Regression model outperforms the other options.

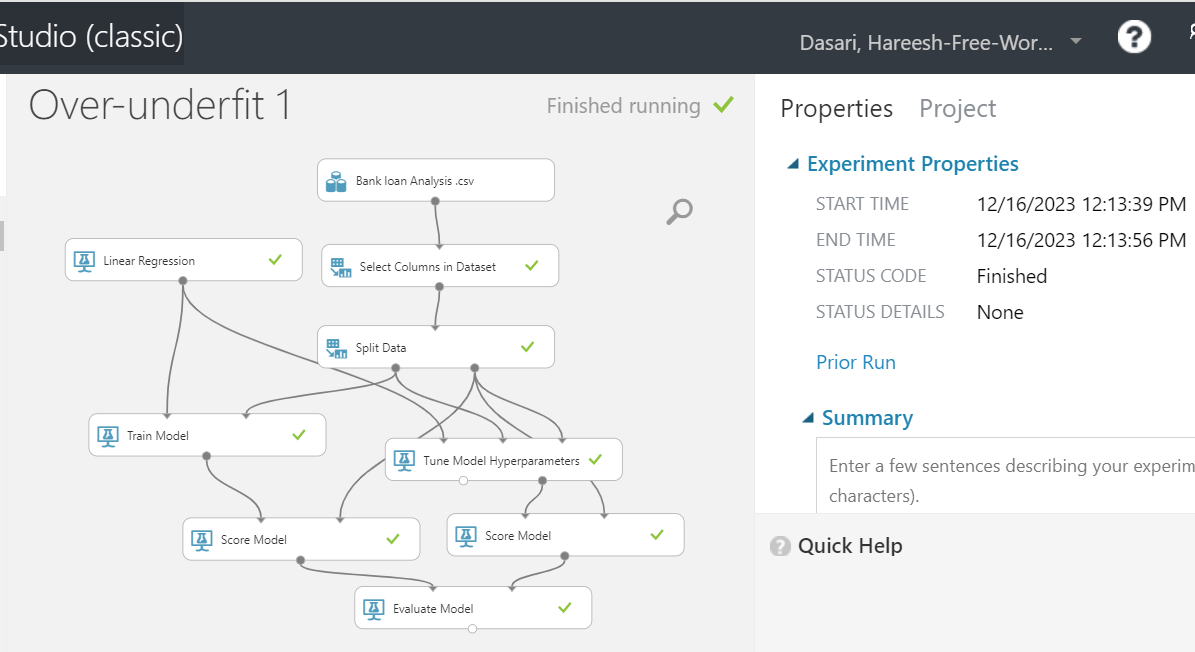
**5 C) Underfitting/ Overfitting Problem:**

Here, we assess the training and test sets for each model. The following validates the testing and training data. Loan Success is the column used as the label column. Linear Regression Model has been used here as the errors are less and CoD is high compared to other two Algorithms.

When a model is too simple to accurately represent the underlying patterns in the data, this happens. High errors in training and poor performance on both training and validation/test datasets are indications of underfitting. Overfitting occurs when a model learns too much from the training data, Low training error but significantly higher error on unseen validation or test data are indicators.

**Models Used for Underfitting / Overfitting Analysis**

**Over-underfit1:**



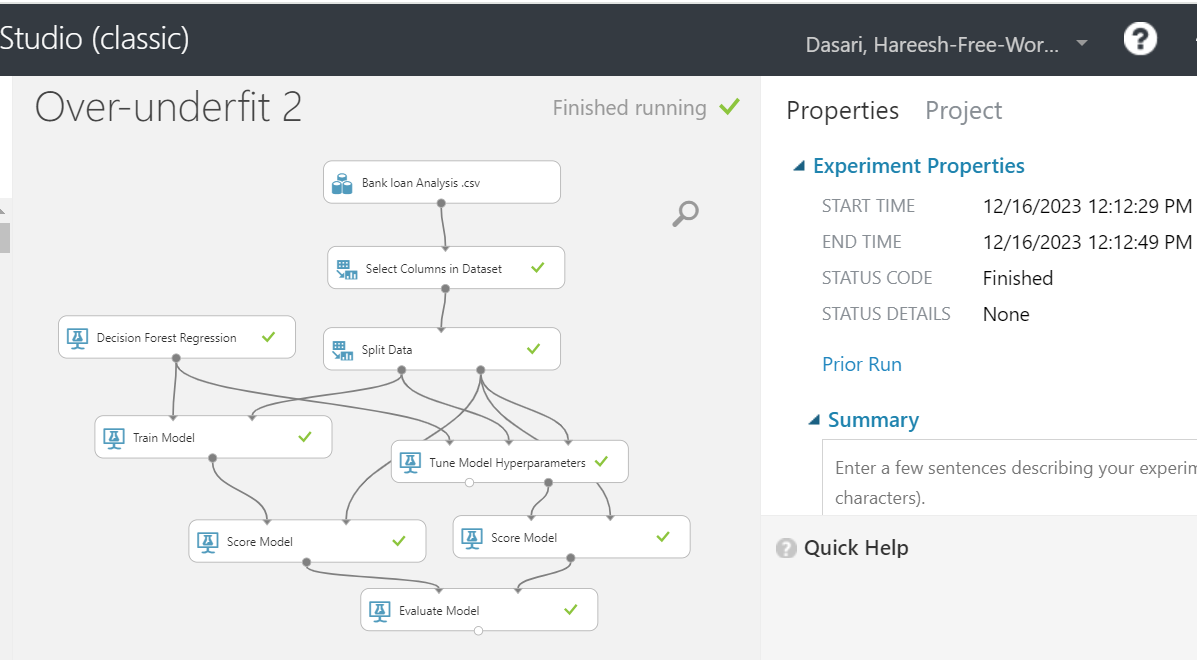
|  |  |  |
| --- | --- | --- |
| **Linear Regression** | **Train** | **Test** |
| Mean Absolute Error | 0.48832 | 0.473414 |
| Root Mean Squared Error | 0.523489 | 0.459547 |
| Relative Absolute Error | 0.979911 | 0.78942 |
| Relative Squared Error | 0.799833 | 0.745839 |
| Coefficient of Determination | 0.59833 | 0.745839 |

The average difference between predicted and actual values is shown by MAE and RMSE values of 0.48832 and 0.523489, respectively. According to the R2 value of 0.59833, the model explains approximately 59.8% of the variance in the training data.

The following test set metrics show improved performance: Lower MAE (0.473414) and RMSE (0.459547) values when compared to the training set indicate better predictive accuracy on unseen data. The higher R2 value of 0.745839 indicates that the model explains approximately 74.6% of the variance in the test data, implying better generalization.

The model performs better on test data than on training data, implying that it generalizes well to new, previously unseen data, displaying improved accuracy and explaining a greater portion of the variability in the test set compared to the training set.

**Model :02**

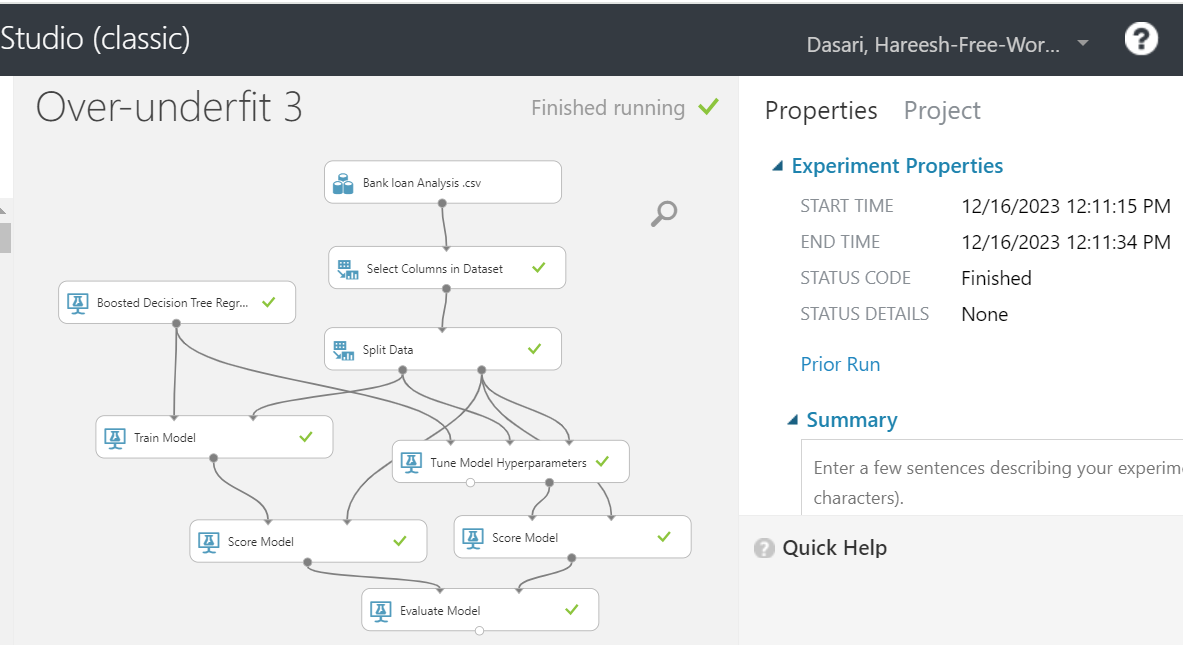


|  |  |  |
| --- | --- | --- |
| **Decision forest Regression** | **Train** | **Test** |
| Mean Absolute Error | 0.402318 | 0.402604 |
| Root Mean Squared Error | 0.415427 | 0.415633 |
| Relative Absolute Error | 0.004638 | 0.605209 |
| Relative Squared Error | 0.762661 | 0.463512 |
| Coefficient of Determination | 0.52661 | 0.653512 |

The training set metrics indicate that model performance is moderate:

The average difference between predicted and actual values is indicated by MAE (0.402318) and RMSE (0.415427). The R2 value of 0.52661 indicates that the model explains 52.7% of the variation in the training data. The metrics from the test set reveal a slightly different scenario: When compared to the training set, similar values for MAE (0.402604) and RMSE (0.415633) indicate consistent performance on new data. The higher R2 value of 0.653512 indicates that the model explains approximately 65.4% of the variability in the test data, indicating better generalization than the training set. The model consistently outperforms on both training and test datasets, demonstrating its ability to generalize well to new, previously unseen data with an improved explanation of the variability in the data.

Model 03



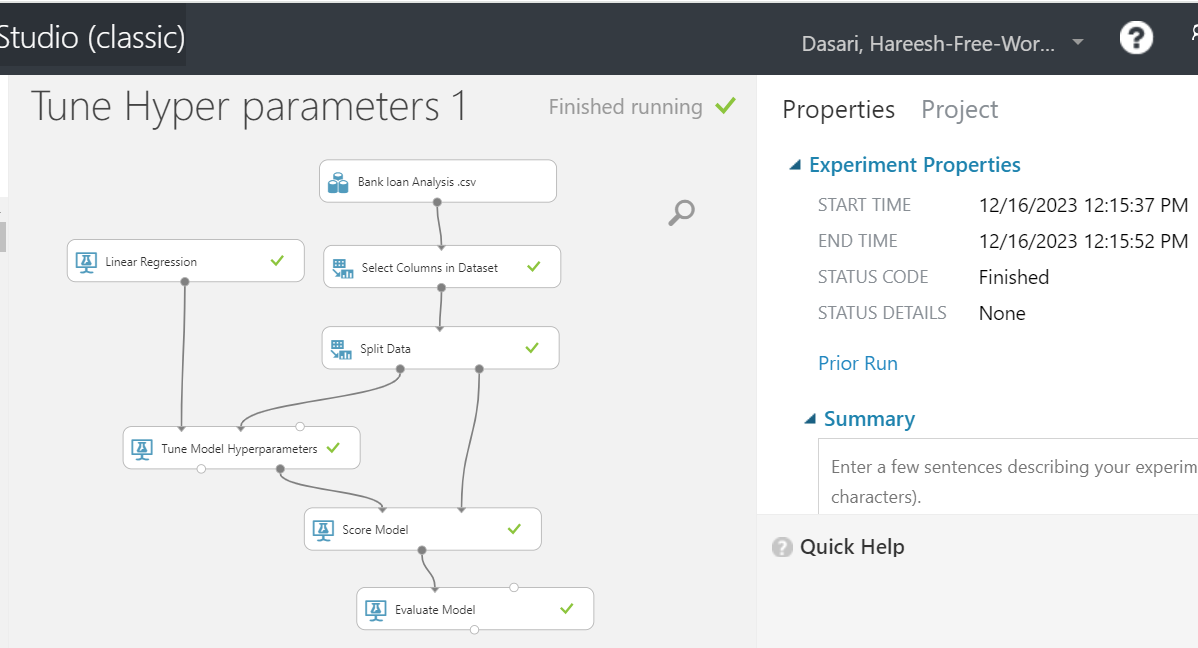
|  |  |  |
| --- | --- | --- |
| **Boosted Decision Tree** | **Train** | **Test** |
| Mean Absolute Error | 0.407417 | 0.496633 |
| Root Mean Squared Error | 0.469796 | 0.542742 |
| Relative Absolute Error | 0.714836 | 0.993267 |
| Relative Squared Error | 0.798674 | 1.178277 |
| Coefficient of Determination | 0.598674 | 0.378277 |

The training data show moderate performance with low errors (MAE: 0.407417, RMSE: 0.469796) and a reasonably high explanatory power (R2: 0.598674). When compared to the training set, test data has higher errors (MAE: 0.496633, RMSE: 0.542742) and lower explanatory capability (R2: 0.378277), indicating weaker generalization and a lower ability to explain variance in unseen observations.

**Conclusion:**  
Linear Regression excels at generalizing to new data, demonstrating consistent improvement from training to test sets, whereas Decision Forest Regression is stable and adaptable to new observations. Boosted Decision Tree, on the other hand, exhibits overfitting tendencies, resulting in lower performance on unseen data despite performing well on the training set.

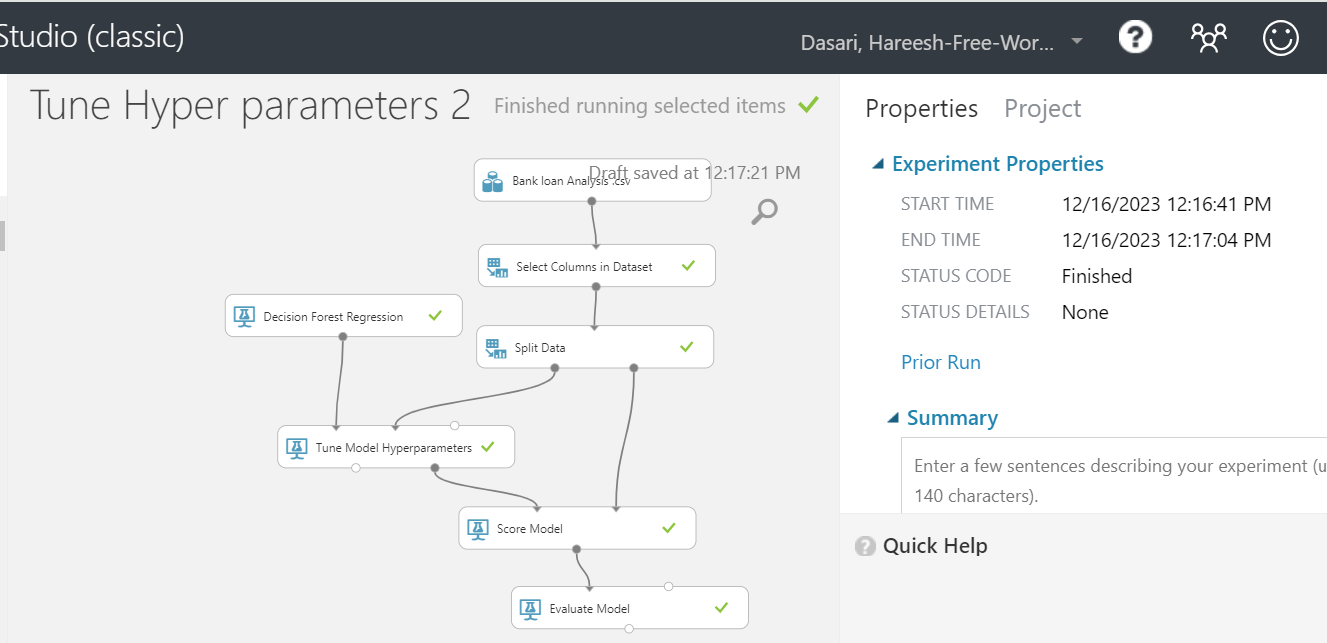
**5 d) Tune Model Hyper Parameters:**

The results of running the model with the Tune Model hyperparameters are displayed below. The model is validated by fine-tuning the performance parameters. The settings for the hyper parameters have been altered and used for this purpose. For the valuation, random sweep and random seed number 899 were used. This process has been repeated for each of the three models in this Modelling process.  
  
**1. Linear Regression:**



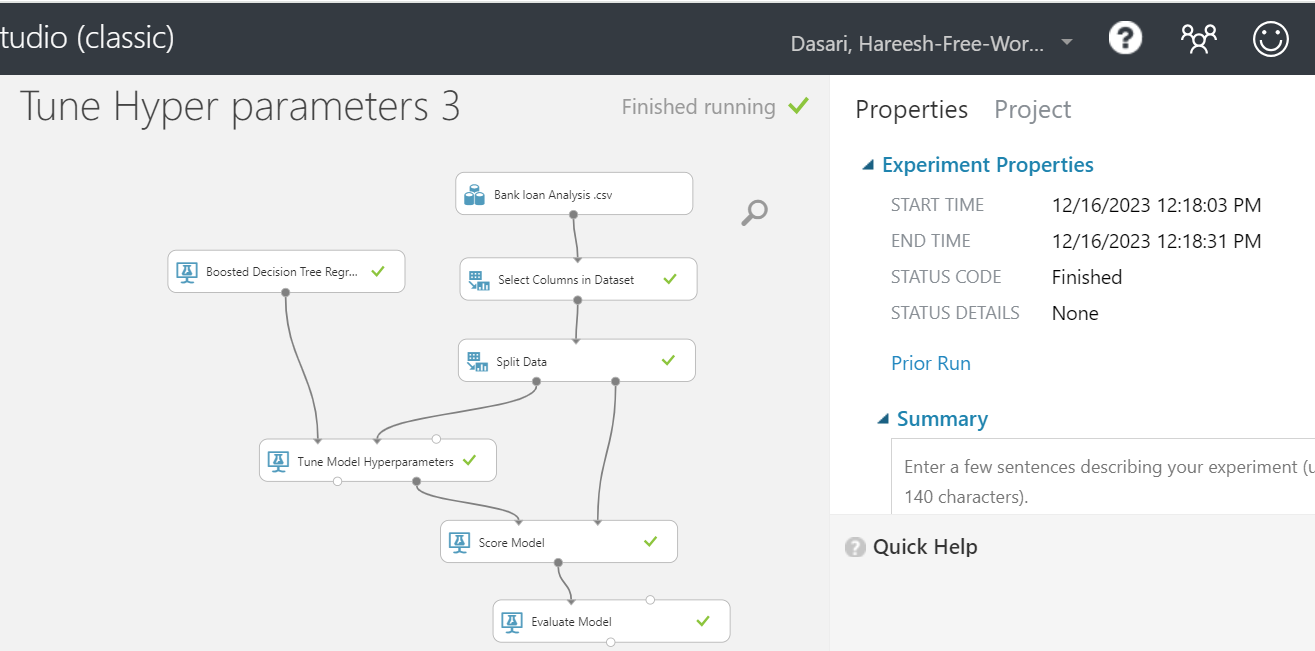
|  |  |
| --- | --- |
| **Parameters** | **Linear Regression** |
|  |
| **Mean Absolute Error** | 0.475069 |  |
| **Root Mean Squared Error** | 0.638958 |  |
| **Relative Absolute Error** | 0.653321 |  |
| **Relative Squared Error** | 0.63854 |  |
| **Coefficient of Determination** | **0.783854** |  |

**2. Decision Forest Regression:**



|  |  |
| --- | --- |
| **Parameters** | **Decision forest** |
|  |
| **Mean Absolute Error** | 0.503727 |  |
| **Root Mean Squared Error** | 0.517762 |  |
| **Relative Absolute Error** | 0.707457 |  |
| **Relative Squared Error** | 0.872312 |  |
| **Coefficient of Determination** | **0.622312** |  |

3. **Boosted Decision Tree Regression:**

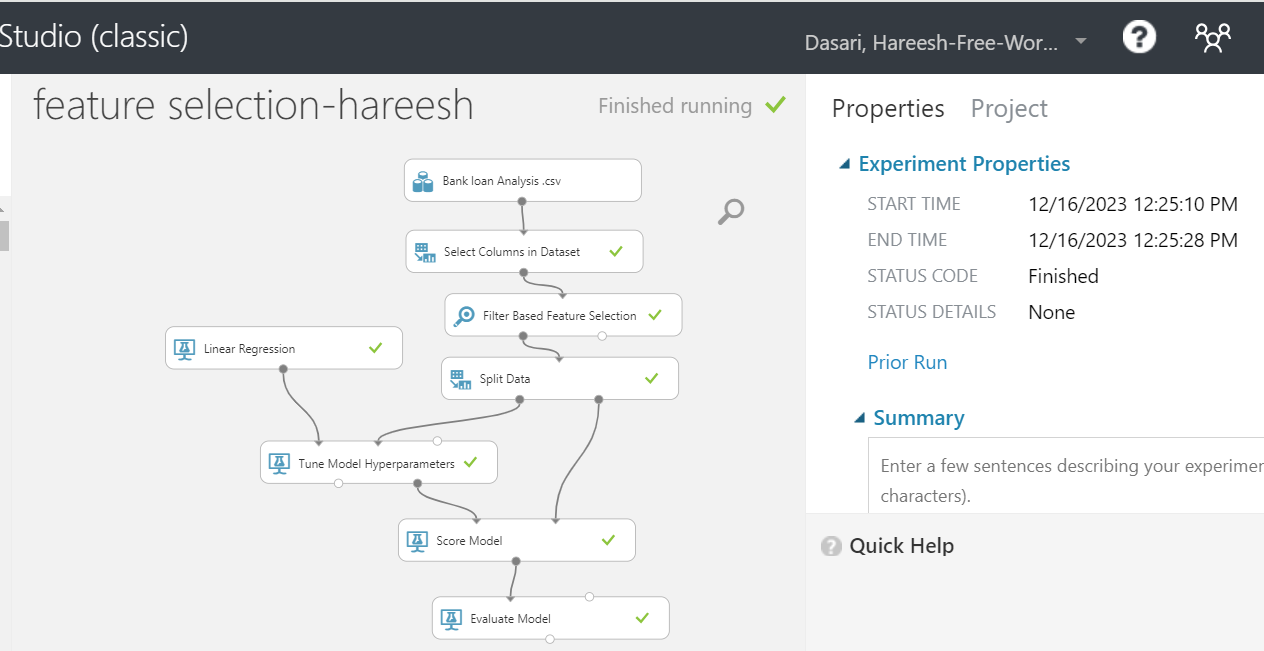


|  |  |
| --- | --- |
| **Parameters** | **Boosted Decision tree** |
|  |
| **Mean Absolute Error** | 0.496633 |  |
| **Root Mean Squared Error** | 0.542742 |  |
| **Relative Absolute Error** | 0.993267 |  |
| **Relative Squared Error** | 0.678277 |  |
| **Coefficient of Determination** | **0.578277** |  |

**Overall Results together:**Linear Regression has the lowest MAE and R2 values of 0.475069 and 0.783854, indicating the lowest average absolute prediction error and the highest proportion of explained variance in the data. When compared to Boosted Decision Tree, Decision Forest Regression has a lower RMSE and higher R2, indicating better overall performance in terms of errors and explanatory power. When compared to the other models, the Boosted Decision Tree has higher error metrics (MAE, RMSE) and lower R2, indicating poorer performance in prediction accuracy and explanatory capability.

**5e) Selecting Features:**

Feature Selection was used here and was applied to all of the models. Following that on all modules, here are the common features discovered for the model. This module specifies the features with the highest predictive power.



Here is the result for features that has the greatest prediction power from this dataset.

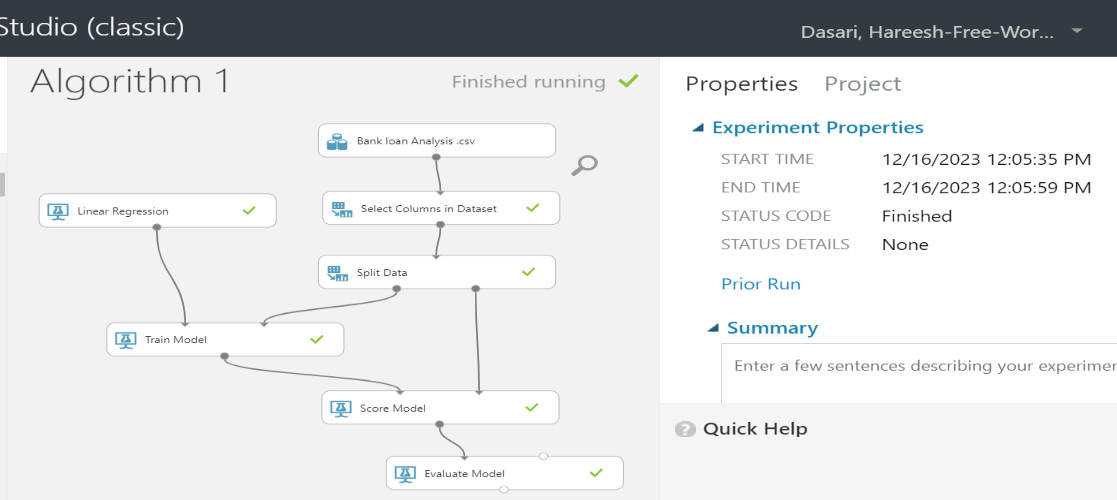
|  |
| --- |
| **Features selection** |
| Family |
| CD Account |
| Income |
| Mortgage |
| Credit Card |
| CC Avg |
| Age |
| Education |
| Loan Success |

These are the above features that has the strong power in predicting the Loans. Feature selection module is used for deriving these features.

**Finalizing The Model:**

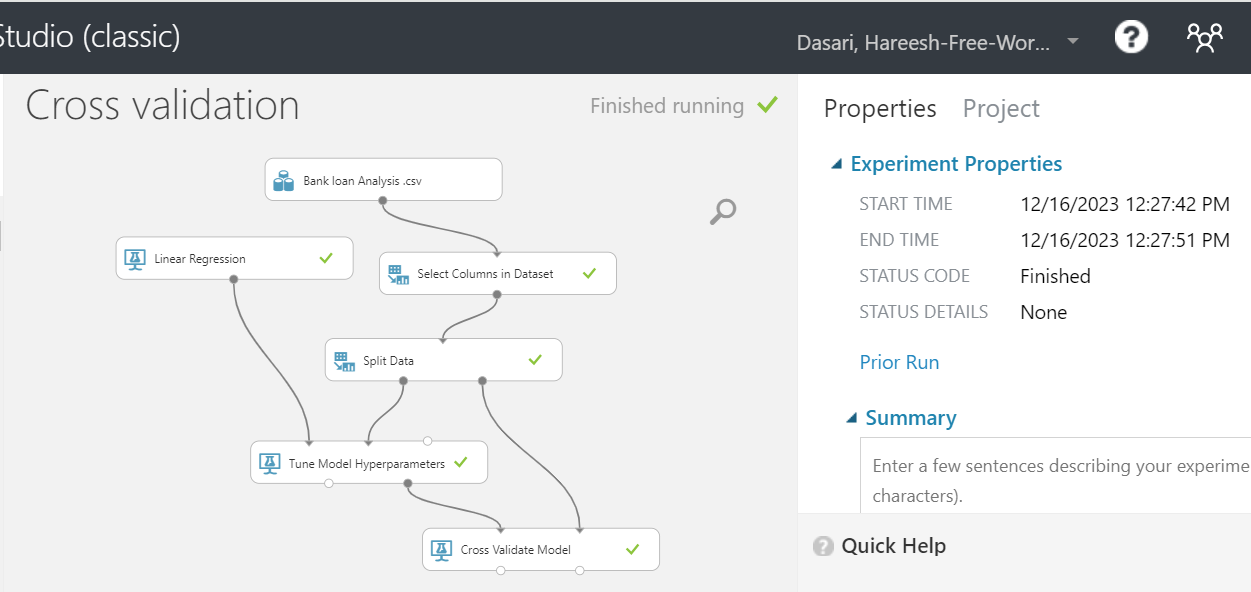
The goal of this research is to create a model for predicting Loan Success. However, it has been determined from the study done so far that Linear Regression has produced better outcomes. And based on the comparison of all the models, this is now finalized. This model has better Coefficient of Determination and less errors compared to other models. Tune Hyperparameters has been used for improvising the metrics. The Cod is 0.743 for this Model.

The finished model is here and is attached below.



**Cross Validation:**

Instead of using Train score and Evaluate model, we utilize Cross validate model. Here, we learn the measures to determine whether it is generating superior outcomes. Here is the model screenshot where I have used cross validate model and captured the results.



**Final Model – Cross Validation Results:**

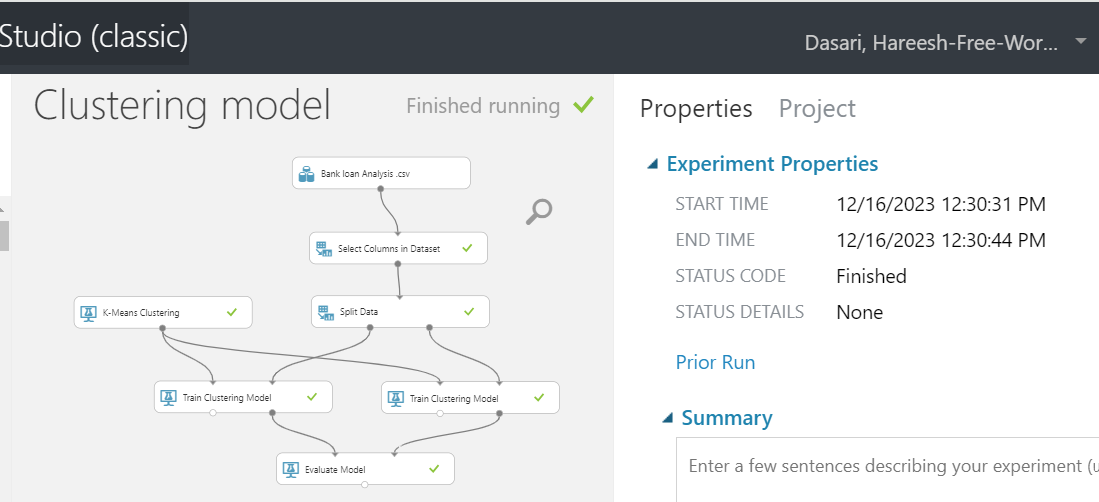
It has been found that the model performs at a rate of 45%. For improved output, the model's tuning is also provided. Mean Absolute error is 0.056 and RMS has reduced to 0.04275 compared to the earlier model. This is because of the tuning and cross validation model which has given the Coefficient of Determination at 45%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mean Absolute Error | Root Mean Squared Error | Relative Absolute Error | Relative Squared  Error | Coefficient of Determination |
| 0.056632 | 0.042785 | 0.113886 | 0.2127 | 0.458225 |

**Training Clustering Model:**

Here, we specify the clusters, and based on the number of columns we're specifying for the model, we perform the training. Here for Here, all of the feature columns are in use, and the settings are shown below.

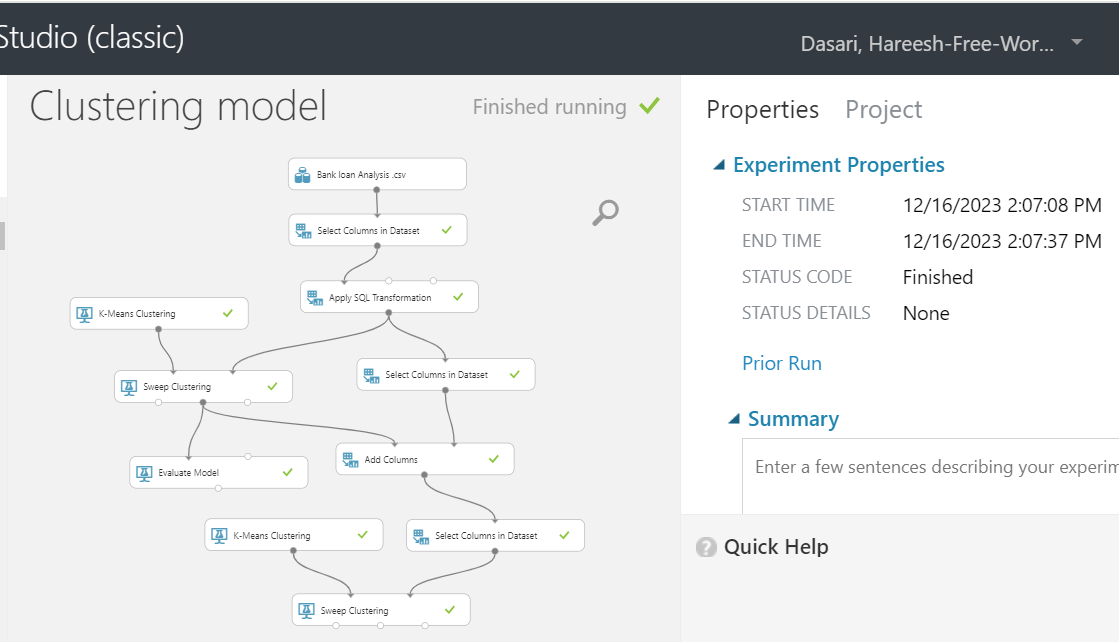
Algorithm K-Means Clustering with default settings

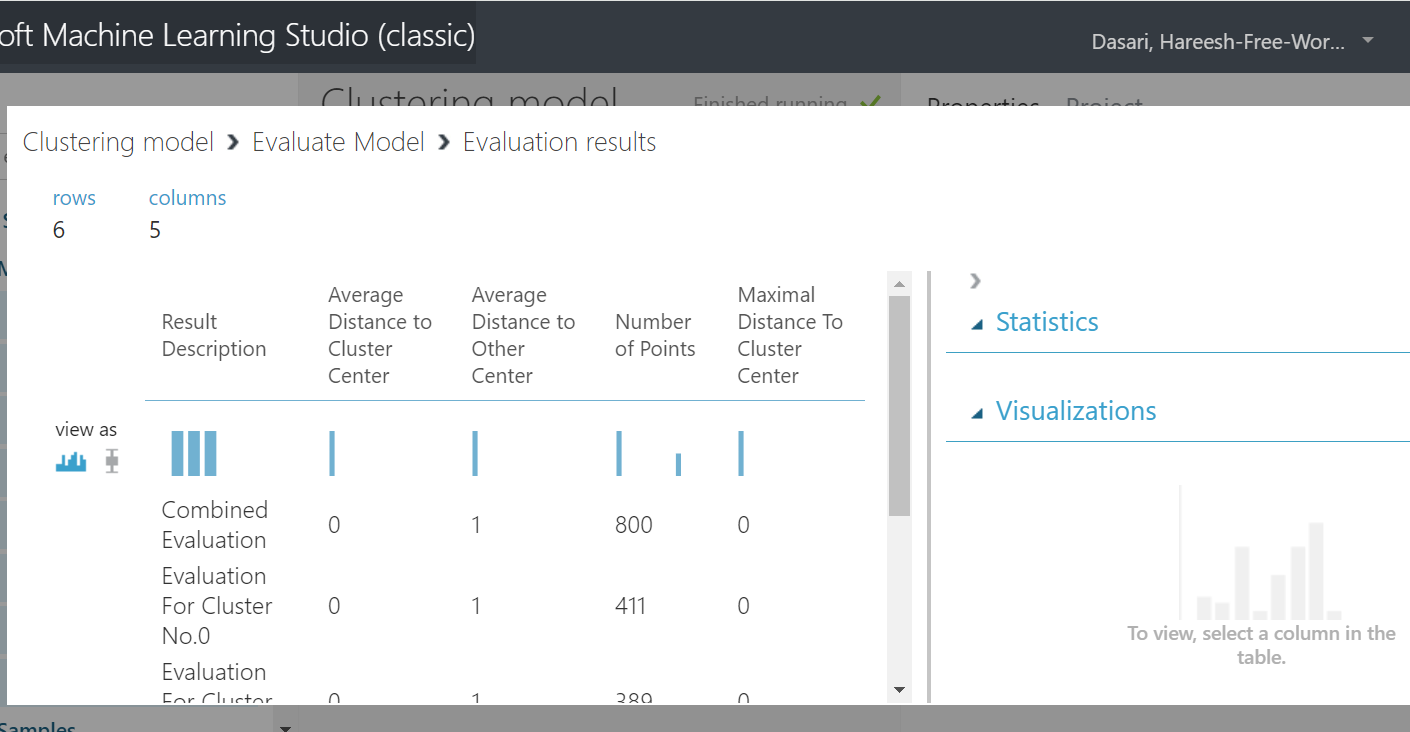


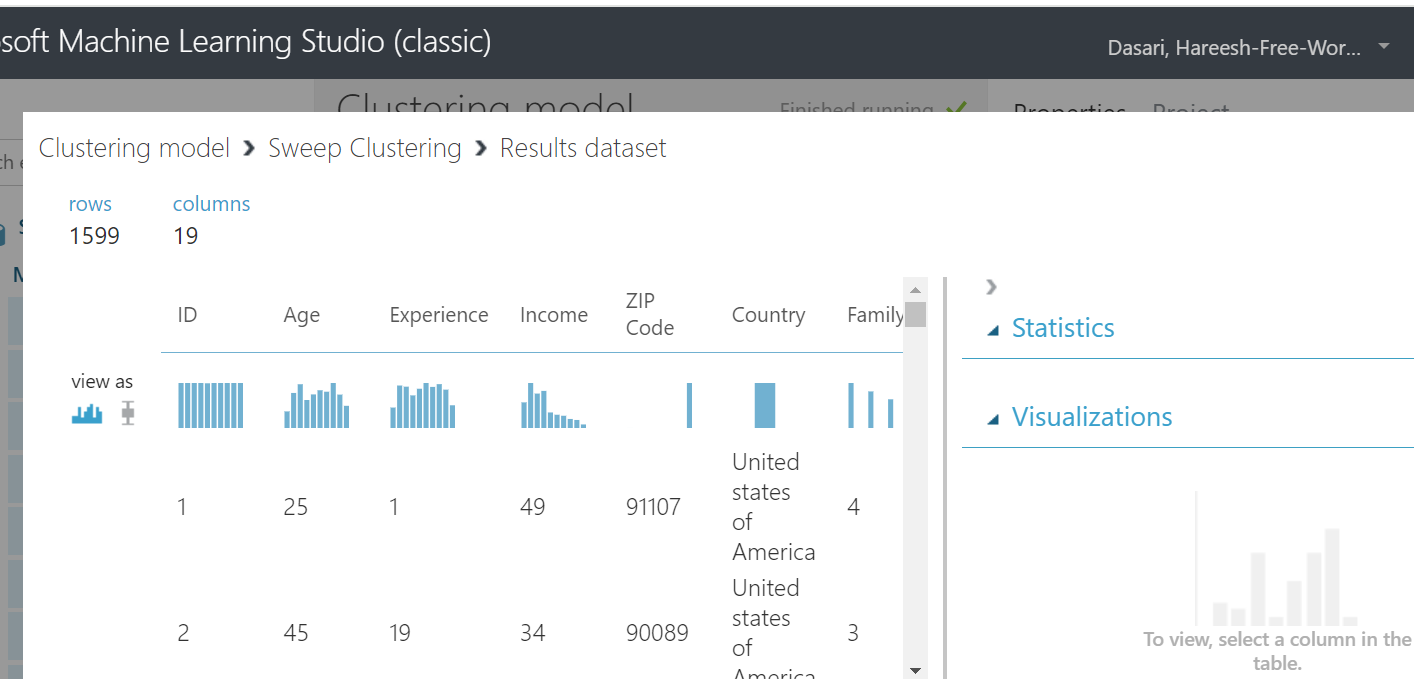
**Clustering with Sweep Cluster and Centroids: Columns Used**

Here the factors that are important for the prediction are the important features like Loan, Income, Age, Marital status are the columns that are used. The problem statement is for defining the loan success based on this parameter.

Here we used Split, Training Data and compared the difference between testing and the training datasets above. Here the centroids are 5 and index of run has varied from 1 till 6, with the cluster metrics being different for all. Attached is the screenshot below for the same.







For the model below, criteria-based clustering is utilized. The current file for homework 7 and this are pretty similar. We have classified the personal loan according to the same guidelines.

**Manual Calculations – Cross Validation:**

According to the instructions, the model is put up here for cross validation. The "Simplified Silhouette" option is used to run the model in the system, and the results are exported to Excel for mean and deviation calculations.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Result Description** | **Average Distance to Cluster Centre** | **Average Distance to Other Centre** | **Number of Points** | **Maximal Distance to Cluster Centre** |
| **Combined Evaluation** | 188.172266 | 749.5998 | 1499 | 374.71942 |
| **Evaluation For Cluster No.0** | 188.262983 | 749.5983 | 750 | 374.71942 |
| **Evaluation For Cluster No.1** | 188.081427 | 749.6013 | 749 | 374.552769 |
| **Mean** | 188.172205 | 749.5998 | 749.5 | 374.636095 |
| **Standard Dev** | **0.128379479** | **0.0021** | **0.707106781** | **0.11784005** |

# **Conclusion:**

The supplied dataset's quality is adequate, but it might be enhanced with the inclusion of fresh data, which would assist in more accurate forecasts. The qualities contribute to the model, according to the model. Because bank loans are so significant, many banks choose to proceed with prudence. As a result, this model offers prediction skills, which has resulted in improved results in terms of finding the best aspects that impact the loan's success rate. Age, salary, education, mortgage, and credit card are among the qualities emphasized. If we apply reasoning to these characteristics, we might claim that this technique provided greater results. This model might be trained using high-end models and datasets to improve the results.