CREDITWORTHINESS

Predicting Default Risk

The Business Problem

I am a loan officer at a small bank operating for two years that needs to come up with an efficient solution to classify new customers on whether they can be approved for a loan or not. Generally my team typically receives 200 loan applications per week and approves them by hand.

Due to a financial scandal that hit a competitive bank last week, I suddenly had an influx of new people applying for loans at our bank instead of the other bank in your city. All of a sudden I had nearly 500 loan applications to process this week. My manager sees this new influx as a great opportunity and wants me to figure out how to process all of these loan applications within one week.

Fortunately I just completed a course in classification modeling and know how to systematically evaluate the creditworthiness of these new loan applicants

Details of project

In this project we will be using credit data to know whether a new loan should get approved or not based on the model we will be training. We will be using 4 models named Logistic Regression, Decision Tree, Random Forest, and Boosted Model and we will test the model and choose the best performed model to predict the creditworthiness of the clients.

Steps performed to get the results

- The very first step I used in the task is to load the data set which includes the creditworthiness of the clients. For that I used an input data tool from the **input tool palette.**
- To know which variable I am going to use and which I am going to deselect from the data, I used a **field summary tool** from a data investigation pallet in which I used all the variables and got the report from the field summary tool. Which helps me to know which variable I am going to use and which variable I am not going to use. After analyzing this report I have deselected 7 variables which I am not going to use in this analysis.
- Variables which I am going to deselect **guarantors, foreign worker, concurrent credits, no.of dependents occupation and telephone** because of lack of variability in the data. Also I am going to remove the **duration_in_current_address** because of missing data.
- Instead of removing the age column as the missing value was only 2% I imputed it with the median.

- After selecting or deselecting the variables in the data now I am going to take the sample of the data to test the model or the training of the model. For that I'm using a **create sample** tool from the **preparation tool palette** in that I am using 70%

estimation sample and 30% for the validation.

- After that I am using **4 classification** models named **logistic regression boosted model decision tree and random forest model**.
- After running these 4 models with an estimation sample I'm going to union the output of all these models using the **union tool** from the **preparation tool palette.**
- Now it's time to compare the models which perform the best. For that I'm using a **model comparison** tool from a **predictive tool** palette and here's the report of model comparison**.**

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression_19	0.7667	0.8444	0.7889	0.9135	0.4348
boosted	0.7400	0.8312	0.7734	0.9231	0.3261
Decision	0.7467	0.8257	0.7340	0.8654	0.4783
forest_model	0.8000	0.8673	0.7606	0.9423	0.4783

Model: model names in the current comparison.

Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

- From the report we can see that the **forest model**, which is a **random forest model**, has given more accuracy in the prediction but we will also look for a **confusing matrix** for a better approach.

Confusion matrix of Decision

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	90	24
Predicted_Non-Creditworthy	14	22

Confusion matrix of Logistic_Regression_19

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	26
Predicted_Non-Creditworthy	9	20

Confusion matrix of boosted

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	96	31
Predicted_Non-Creditworthy	8	15

Confusion matrix of forest_model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	98	24
Predicted_Non-Creditworthy	6	22

- Decision tree has given 74% of accuracy which is the average accuracy of the other two models which is logistic regression and decision tree. But this model seems to be biased because positive predicted values are 75% and negative

predicted values are 61%. The difference of 14% is huge.

- Confusion matrix of Decision

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	90	24
Predicted_Non-Creditworthy	14	22

- Logistic regression has the accuracy of 76% and if we talk about confusing metrics the positive predicted values are 79% and negative predicted values are 61% the difference of 18% is huge the model seems to baised.

- Confusion matrix of Logistic_Regression_19

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	26
Predicted_Non-Creditworthy	9	20

- Boosted model has the accuracy of 74% and positive predicted values 75% and negative values 65%. The difference of 10% is also huge.

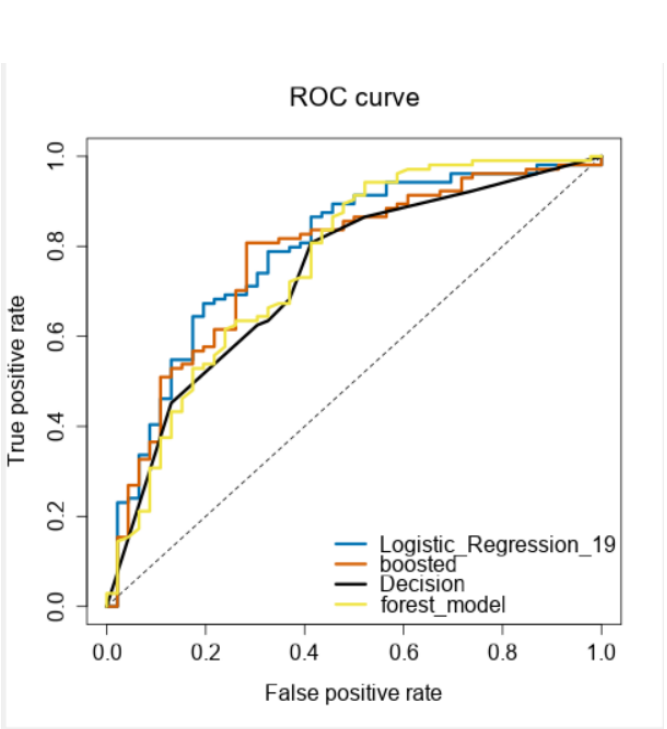
- Confusion matrix of boosted

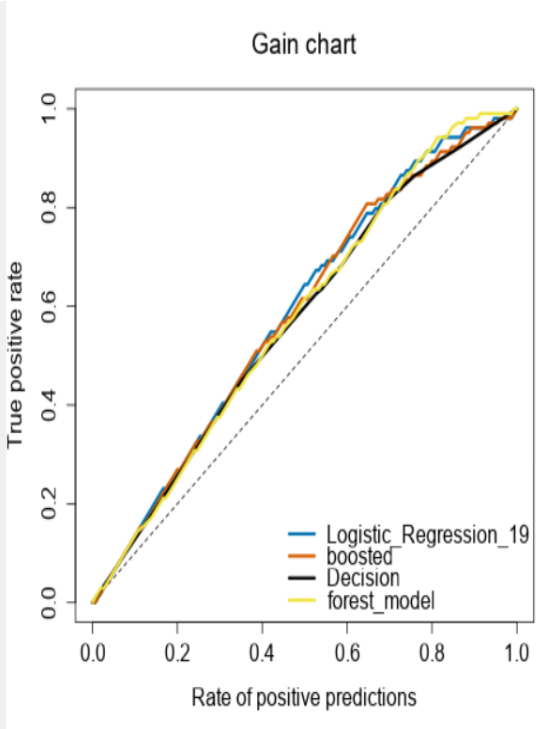
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	96	31
Predicted_Non-Creditworthy	8	15

- Random forest model has produced the highest accuracy at 80% and the positive predicted values 85% and negative predicted values are 78% the differences only 7% which is the lowest so we can say the best model is random forest model for this analysis.

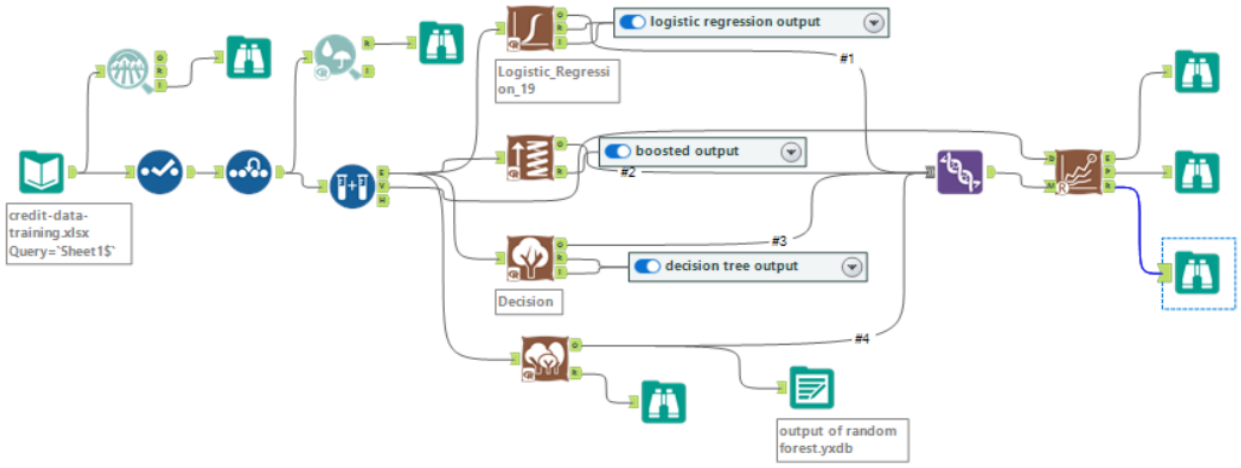
Confusion matrix of forest_model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	98	24
Predicted_Non-Creditworthy	6	22

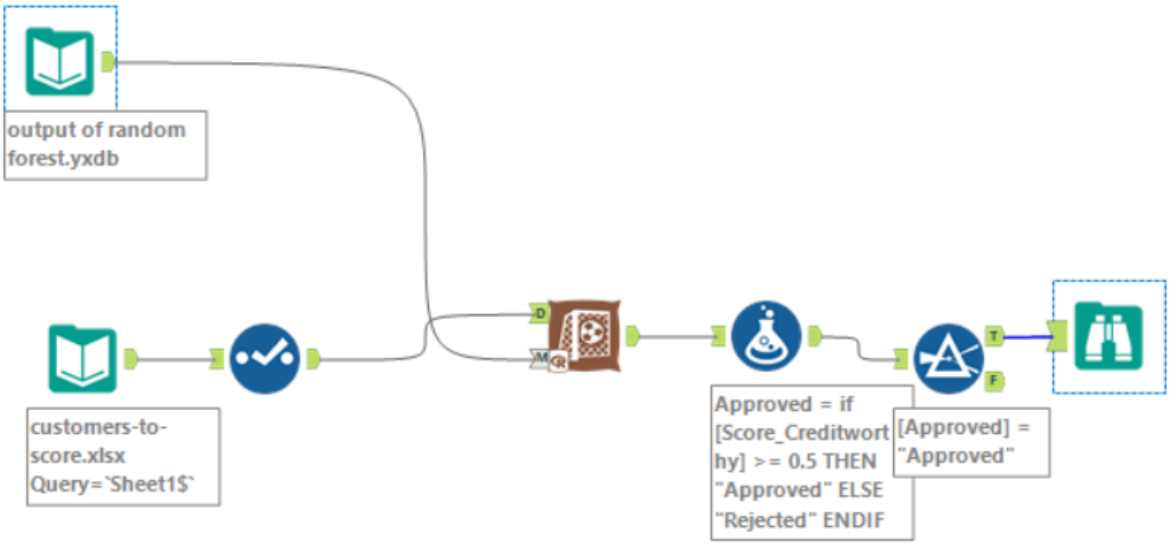




Now here I am attaching the whole workflow of the ALTERYX.



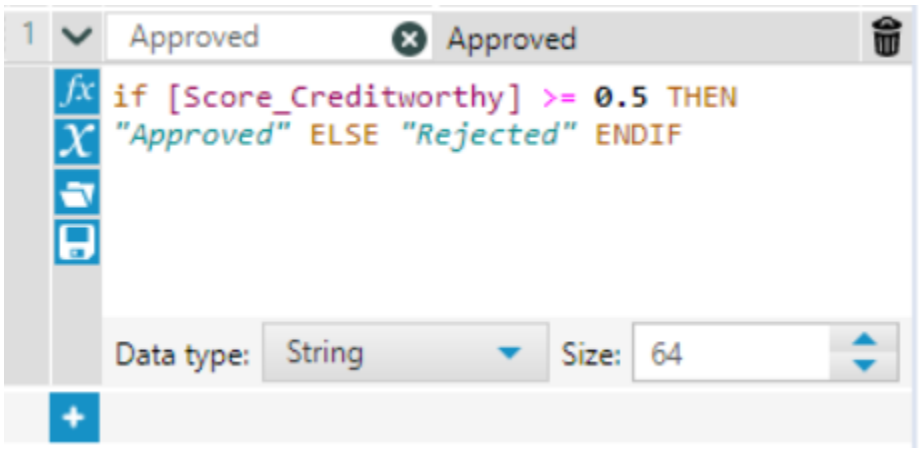
This part of work flow is related to the model training and model comparison where we trained the model and compared the model with the sample data which we have created from the total data set we had.



This is the work flow where we predicted the values using a model which we have created in the prior workflow and now we are scoring the data using a **score tool** from the predictive tool palette.

After giving scores we used a **formula tool** to get the approved or rejected values

from the score.



and then we used a filter tool to filter out the proved results. The total approved results are **414** that means we have approved 414 loan applications out of 500.