

**Project Report**

**“BigMart Sales Prediction”**

**Phase 2: (Model Training)**

**Name:** Mohammad Hunain | **Course:** Machine Learning | **Date:** 23-Dec-2018

TABLE OF CONTENTS:

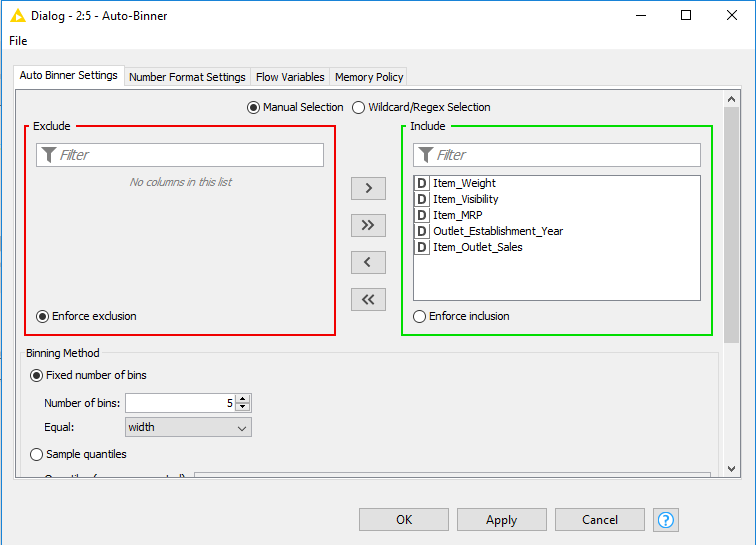
1. Abstract 2
2. Binning & Portioning 3
3. Training My First Model 4
4. Machine Learning Algorithm 4
5. Decision Tree (Information based learning) 4
6. Prediction 5
7. ROC Curve 6
8. Confusion Matrix 7
9. K-Nearest Neighbor (Similarity based learning) 8
   1. Predicted Data 8
   2. ROC Curve 9
   3. Confusion Matrix 10
10. Naïve Bayes (Probability based learning) 11
    1. Predicted Data 11
    2. ROC Curve 12
    3. Confusion Matrix 13
11. Linear Regression(Error based learning) 14
    1. Predicted Data 14
    2. ROC Curve 15
    3. Confusion Matrix 16
12. Results 1790
13. Final Workflow 17

**ABSTRACT**

This report describes a project performed in machine learning techniques and analytics tools. The aim of the project was the predicting the business problem through the Data set of BigMart Sales and discussing the goals of business problem. The main focus of this project is to increase the sales of any SuperMarts or BigMarts”. So the idea is to find out the properties of a product, and store which impacts the sales of a product.

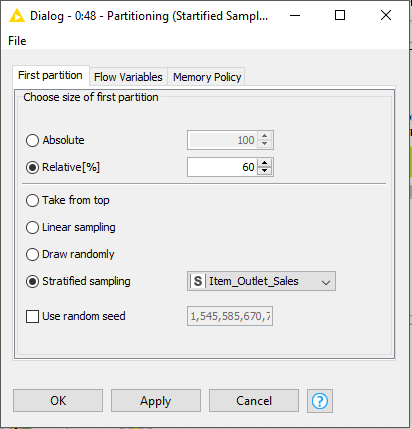
This Report is only the phase 2 of my project. After Successfully Evaluation and preprocessing the Data, in this Phase I train my first ML Model with four Machine Learning Algorithms.

**Auto BINNING:**



**PARTITIONING:**

Stratified Sampling (60:40)



**1. Training My First Model**

The process of training an ML model involves providing an ML algorithm (that is, the learning algorithm) with training data to learn from. The term ML model refers to the model artifact that is created by the training process.

**2. Machine Learning Algorithms**

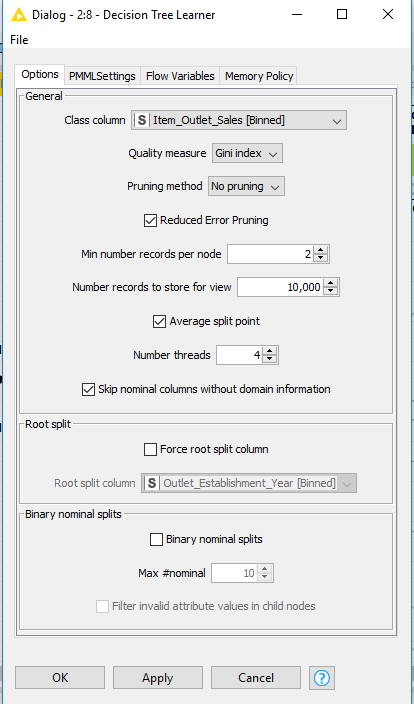
To get a prediction, I used four machine learning algorithms.

* Information Gain
* Similarity Based
* Probability Based
* Error Based.
* **Information Base Learning**

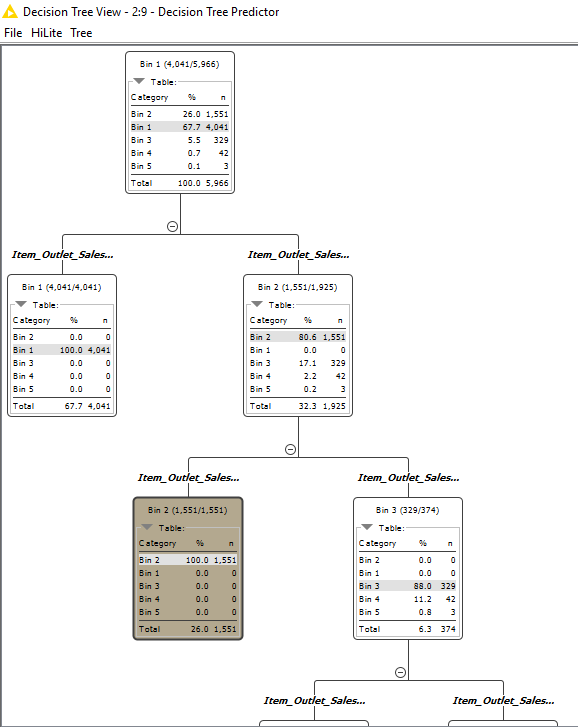
Decision Trees are an important type of algorithm for predictive modeling machine learning.

The representation of the decision tree model is a binary tree. This is your binary tree from algorithms and data structures, nothing too fancy. Each node represents a single input variable (x) and a split point on that variable (assuming the variable is numeric).

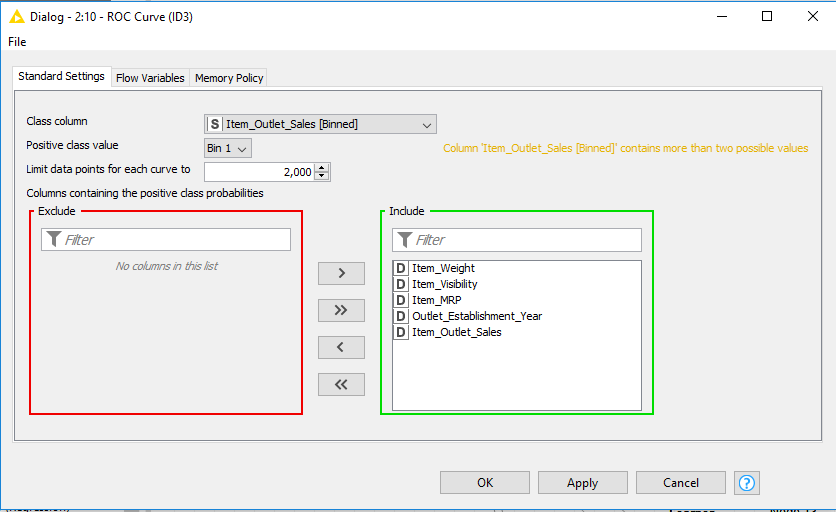
* **Configuration:**

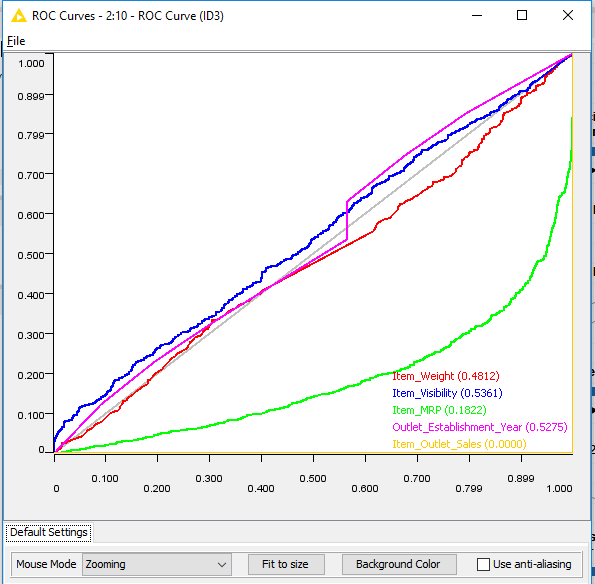


* **Decision Tree Predictor:**

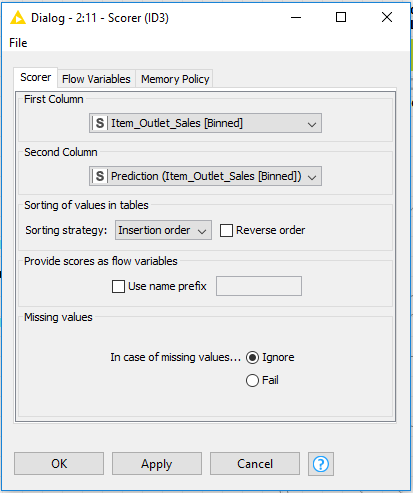


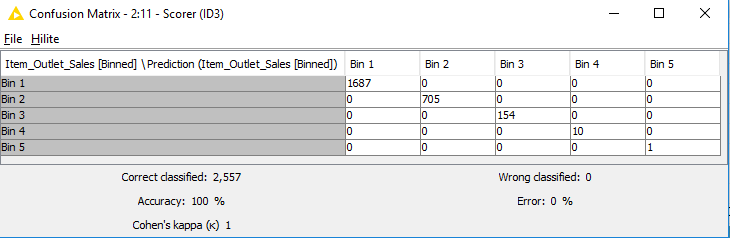
* **ROC Curve of Decision Tree**





* **Confusion Matrix of Decision Tree**





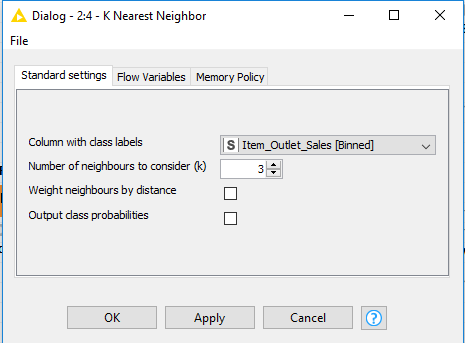
2nd Algorithm

* **Similarity Based Algorithms**

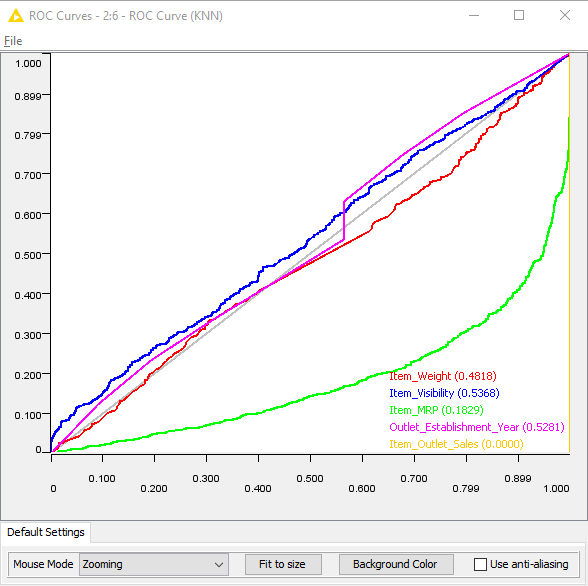
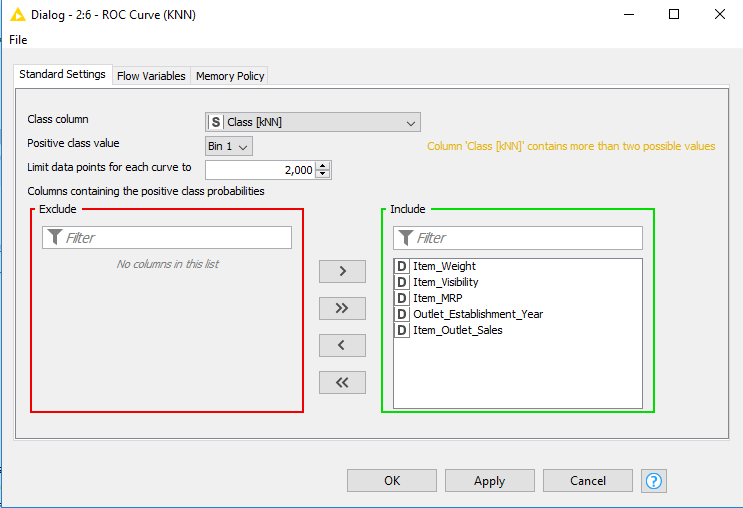
In Similarity Based Learning we use KNN algorithm which is very simple and very effective. The model representation for KNN is the entire training dataset. Simple right?

Predictions are made for a new data point by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those K instances. For regression problems, this might be the mean output variable, for classification problems this might be the mode (or most common) class value.

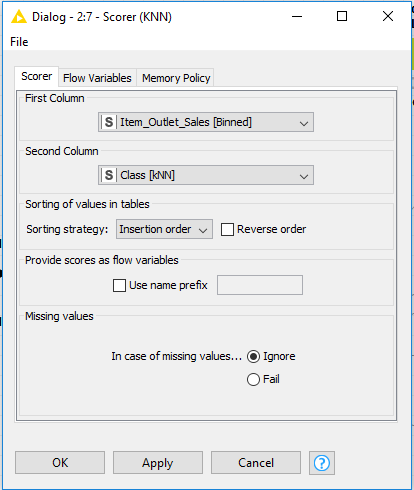
* **Configure**

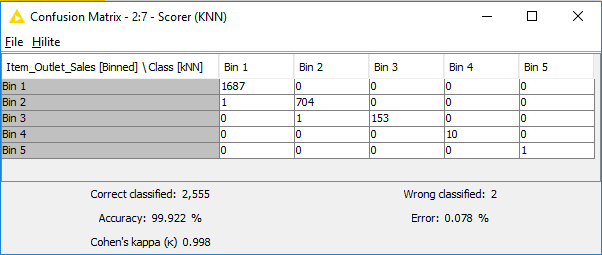


* **ROC curve of K-Nearest Neighbor**



* **Confusion Matrix of K-Nearest Neighbor**





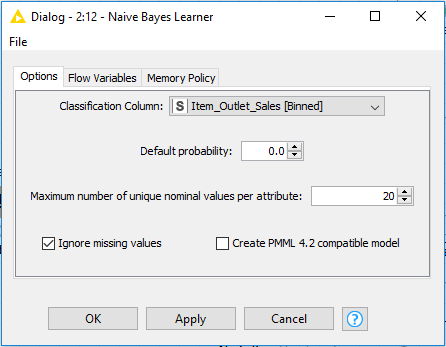
3rd Algorithm

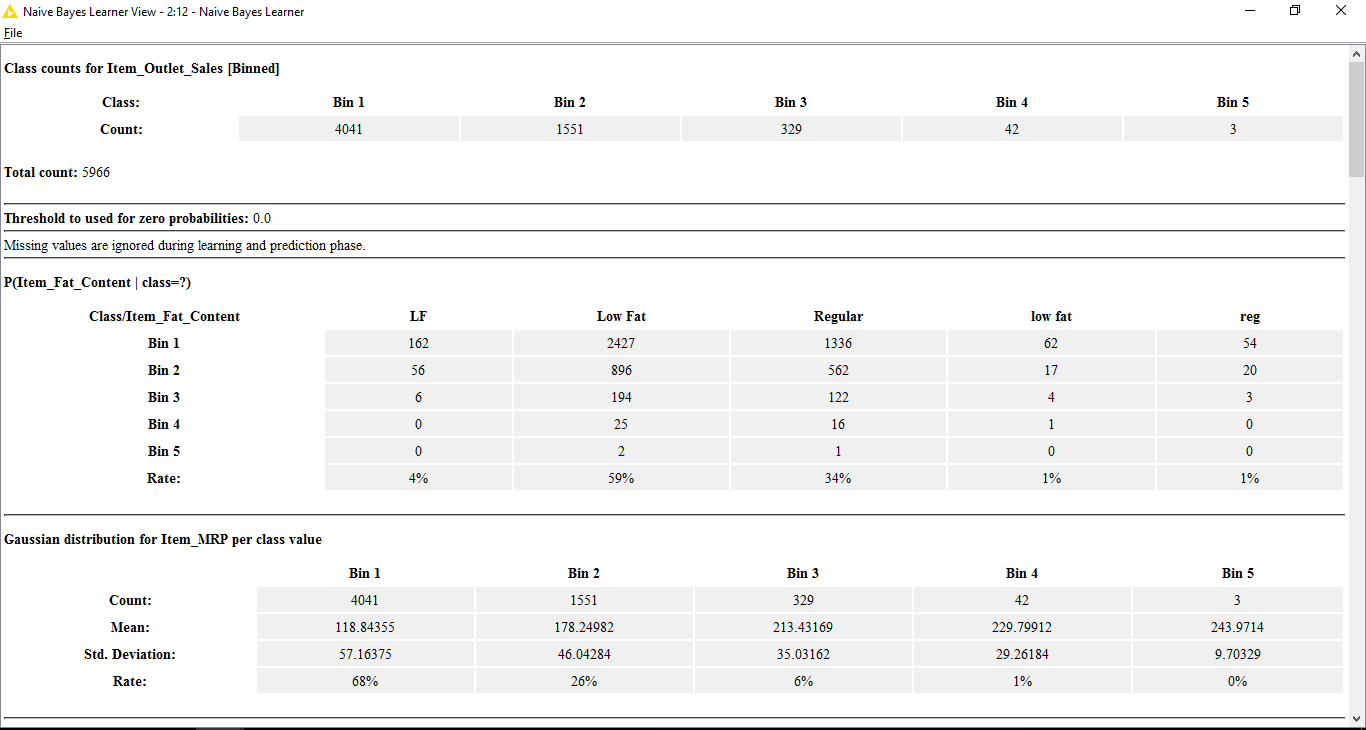
* **Probability Based Algorithms**

Naive Bayes is a simple but surprisingly powerful algorithm for predictive modeling.

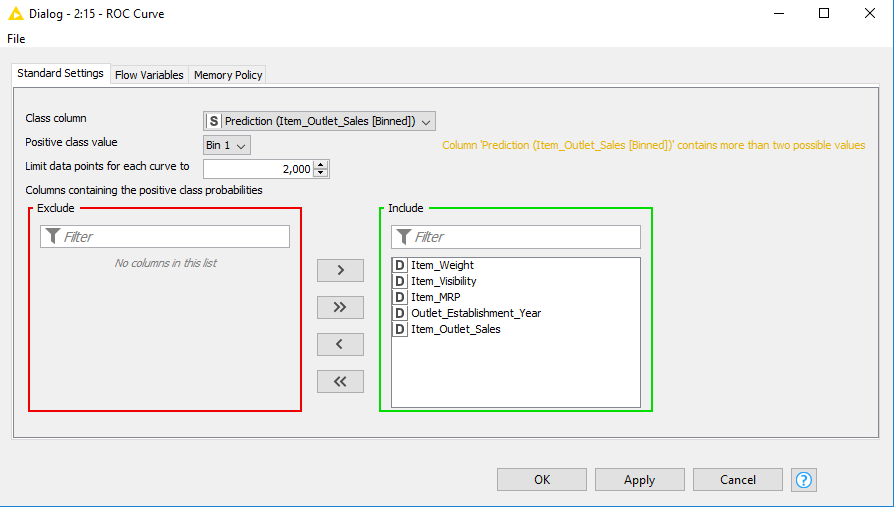
The model is comprised of two types of probabilities that can be calculated directly from your training data: 1) The probability of each class; and 2) The conditional probability for each class given each x value. Once calculated, the probability model can be used to make predictions for new data using Bayes Theorem.

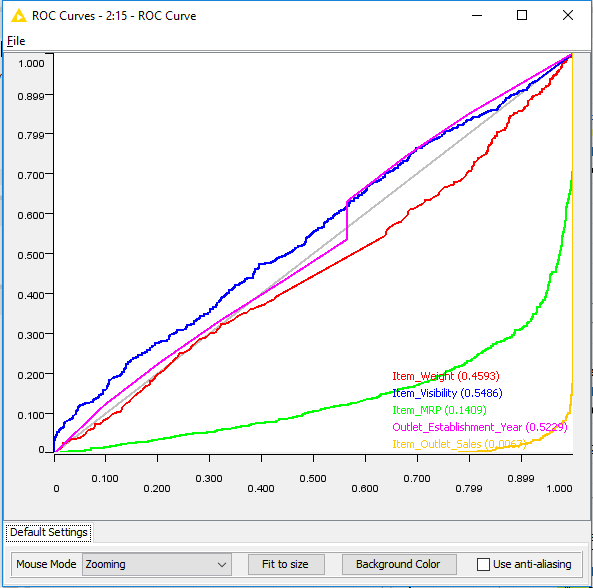
* **Configure**



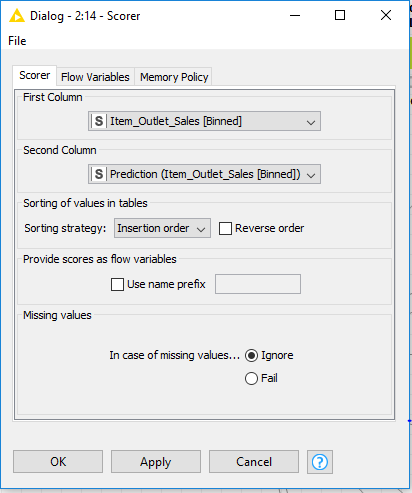


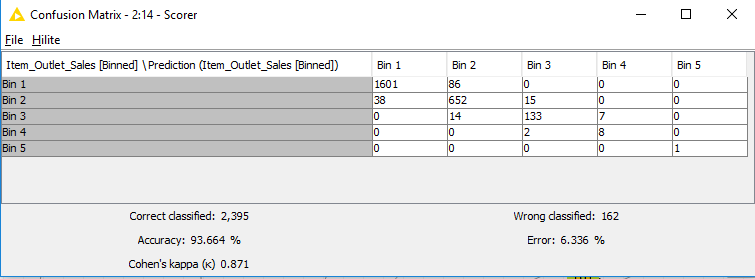
* **ROC of Naive Bayes**





* **Confusion Matrix of Naive Bayes**



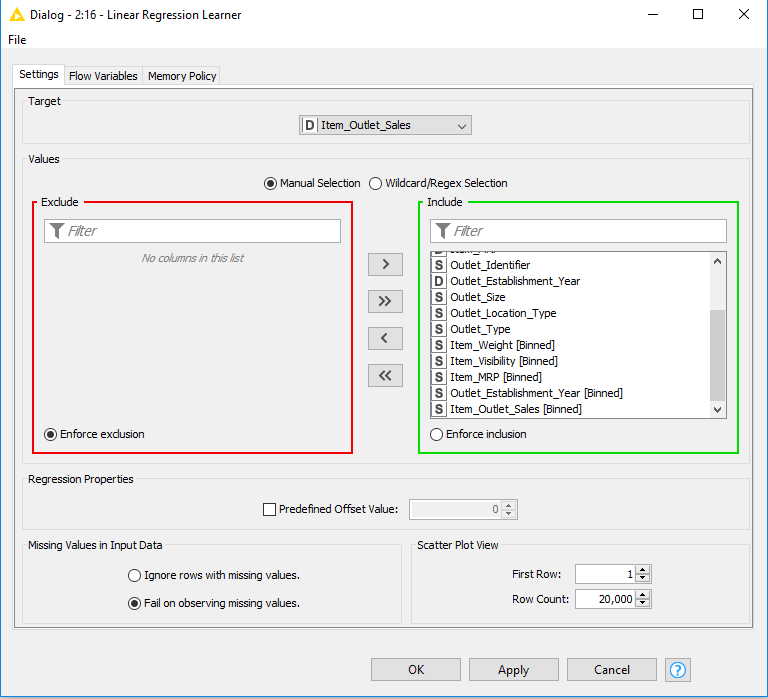


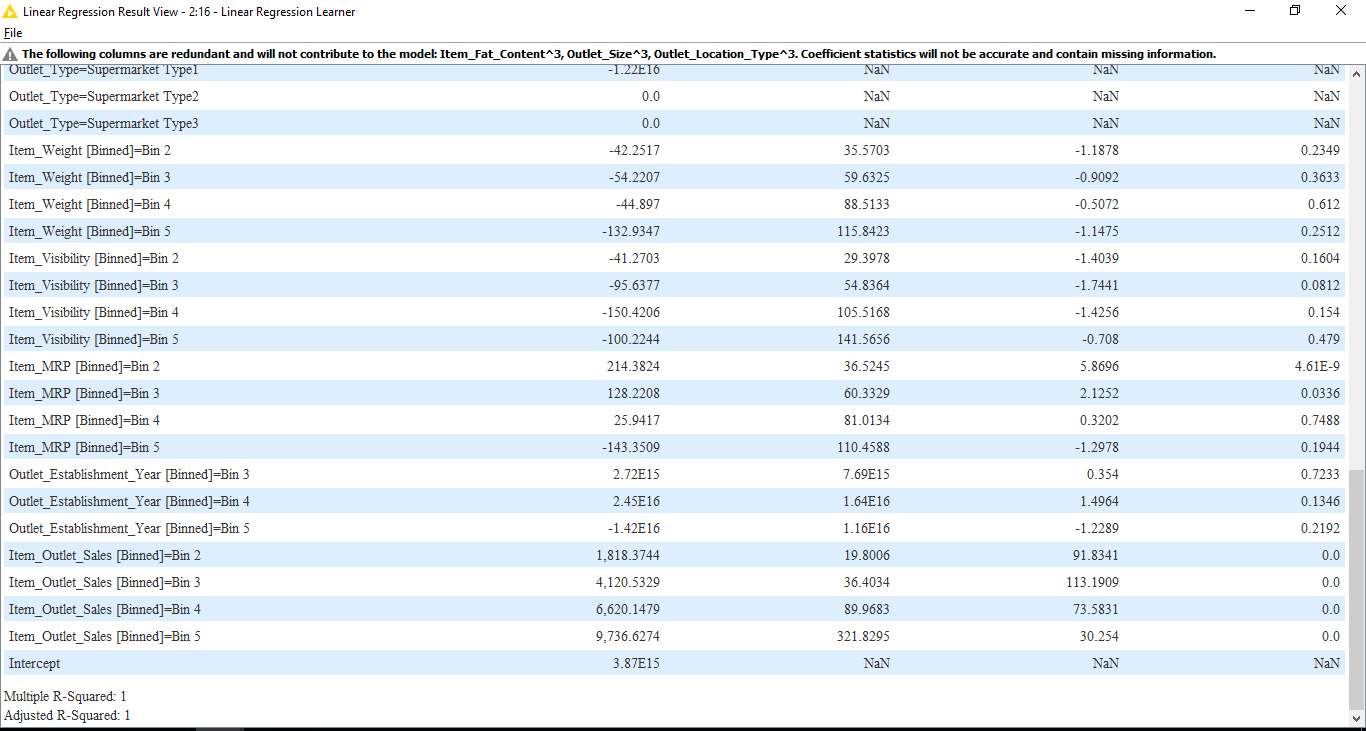
4th Algorithm

* **Error Based Learning**

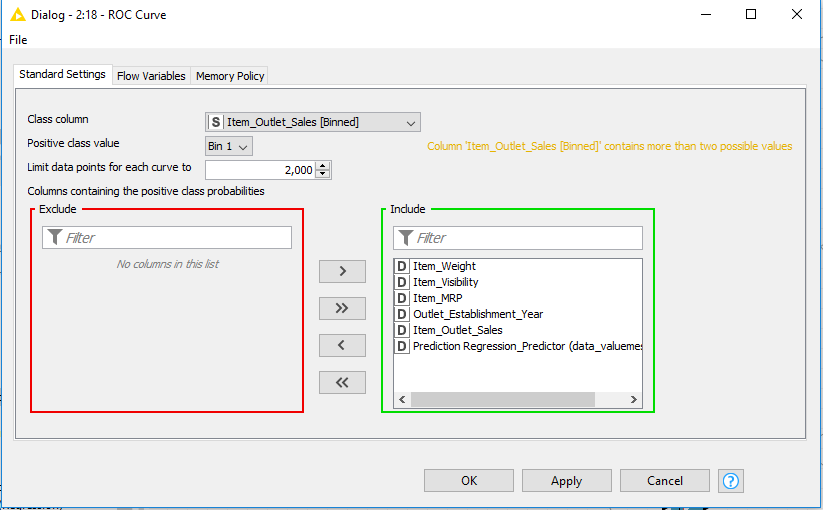
Linear regression is perhaps one of the most well-known and well-understood algorithms in statistics and machine learning. Predictive modeling is primarily concerned with minimizing the error of a model or making the most accurate predictions possible, at the expense of explain ability. We will borrow, reuse and steal algorithms from many different fields, including statistics and use them towards these ends.

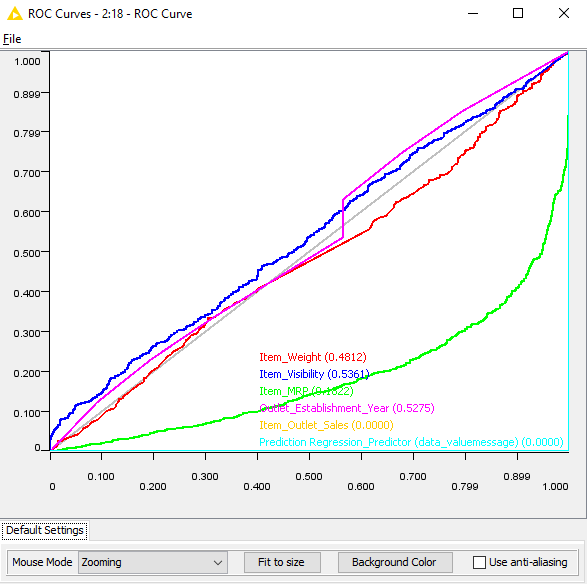
* **Configure**



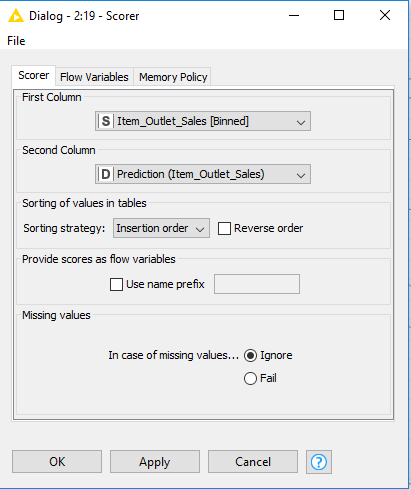


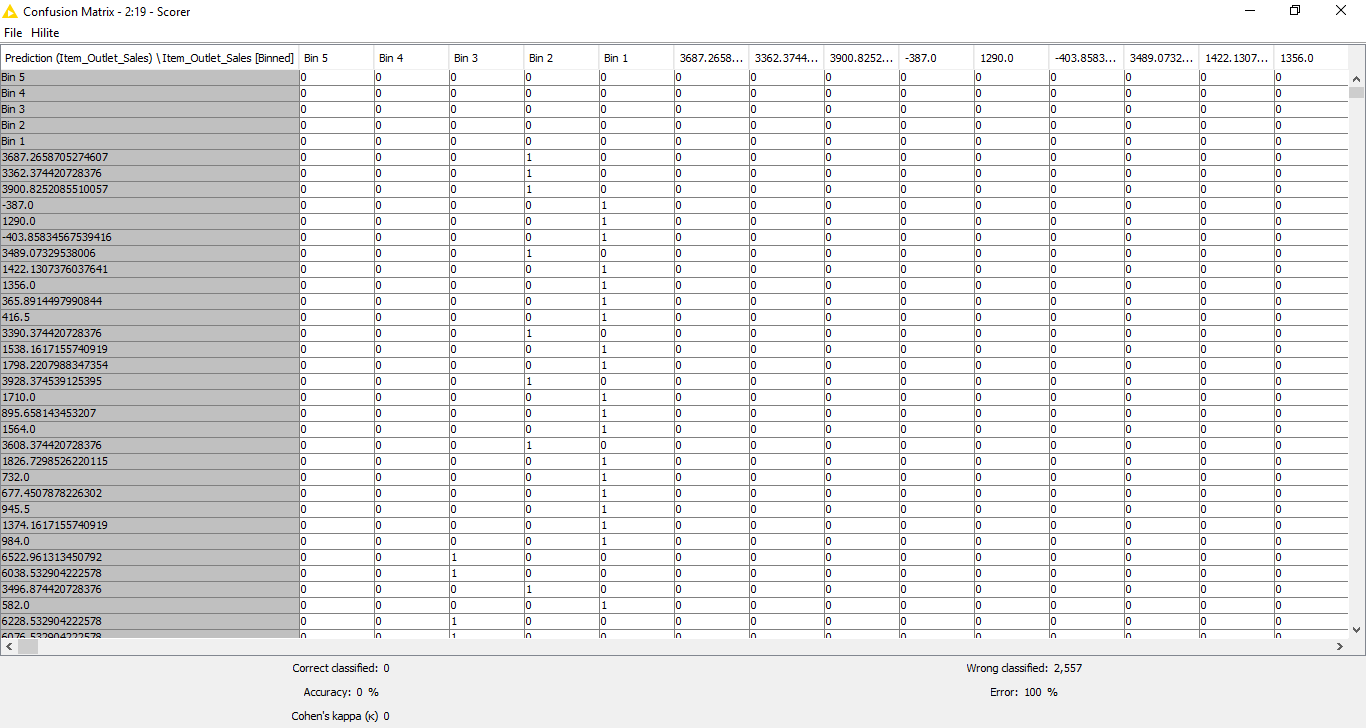
* **ROC of Linear Regression**





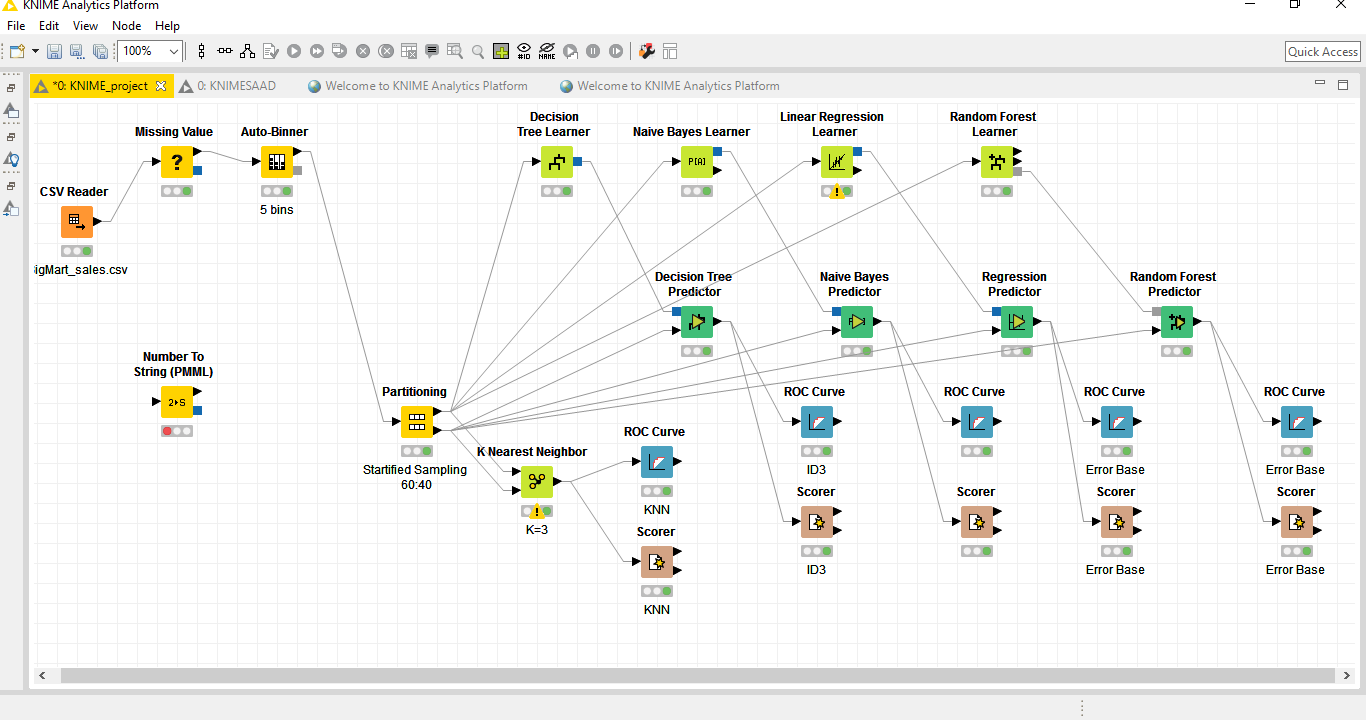
* **Confusion Matrix of Linear Regression**





# **Results**

Out of all four algorithms applied, Decision Tree of Information Gain and KNN of Similarity Based Learning gave the best prediction and 0.00 ROC Curve with Accuracy of above 90%. Hence, these two performed the best with the dataset of the above-mentioned business problem.

**FINAL WORKFLOW**