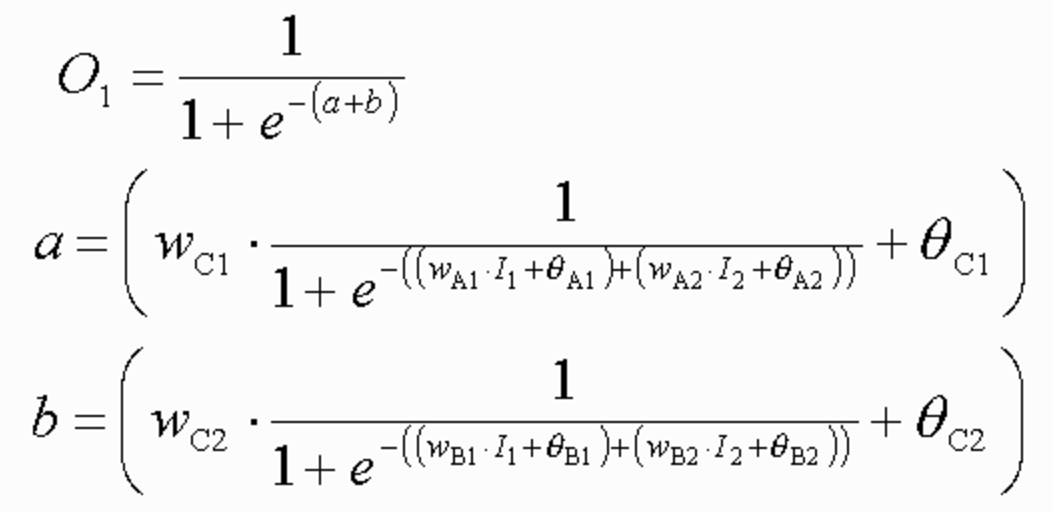
**Capstone Project: Weather affect food sales?**

**Contribution of Competitor article:**

Weather is major factor for food sales the food consumed by the people various drastically based on the type of weather/season the food is grown and the taste of the food, The food is also imported from a different location in usual causes and the transportation also various based on the speed of the mode. My competitor used neural network with 2 nodes to identify the maximum accuracy.



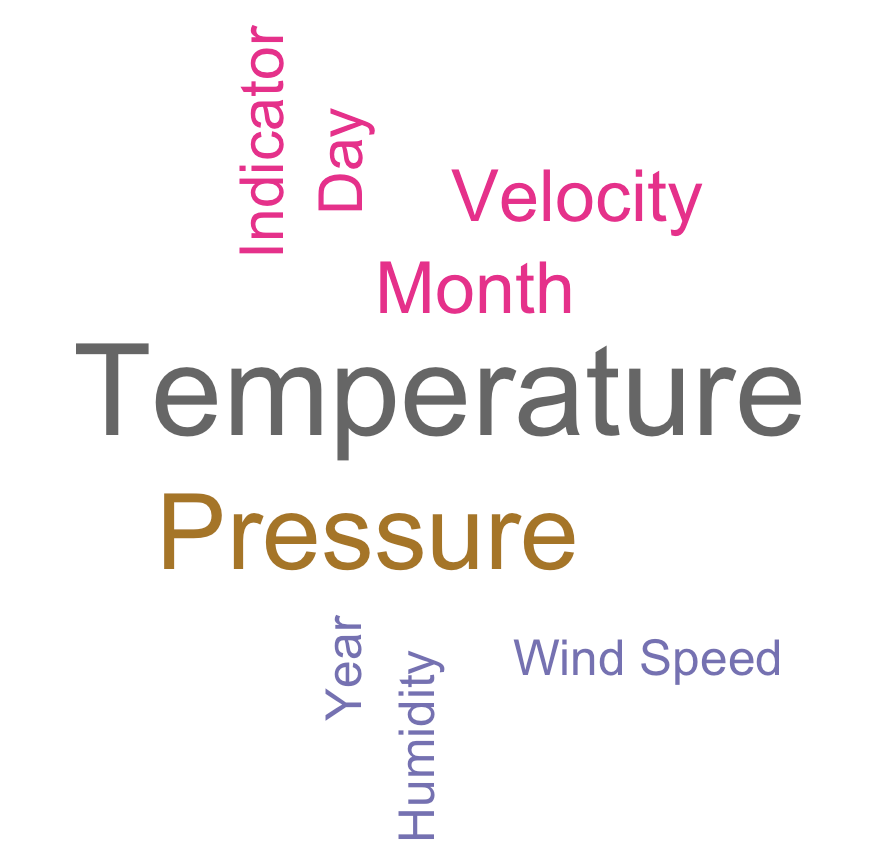
The competitor used the methods carefully and designed the techniques in such a way that the system gives the best result over the course and the model doesn’t over fit the requirement. But to out perform the methodology used by my competitor I have improvised the ETL (Extract Transfer Load) to the optimized situation that it pulls all the data periodically and displays the result real time situation, Spotfire has been used as a dashboard tool to visualize the graph real time to provide filters for the variables (independent variables) and summarize the results to the end users to simplify the process and act to the needs.

**Description of my Contribution:**

* Used modern ETL tools to increase the performance
* Interactive live dashboards for better data interpretation.
* Used simpler methodology to produce the same result as my competitor.
* Designed/Structured the methodology to work with the like data and used TERR to data manipulate using Spotfire.

**Data Source and Content:**

Weather Data set:



Sales Data set:

****

The data consists of the following columns

1. Temperature
2. Humidity
3. Pressure
4. Dew point
5. Wind speed
6. Wind direction
7. Visibility
8. Timestamp

Let’s look at the description of all these variables in depth below.

Temperature :

Temperature is a measure used to identify if the surrounding is hot or cold.

Humidity :

It tells the moisture content of the atmosphere. When the humidity is high it feels oppressive outside because of sweat doesn’t evaporate. When the humidity is low you feel cooler but moisture evaporated. The relative humidity is dependent on air temperature. If the water vapour content stays the same and the temperature drops the relative humidity increases.

Pressure :

Air gets lighter with increasing altitude.

Dew Point:

Dew point may drop by some possible influences; as when the temperature drops it’s enough to get fog and dew, water condenses out of the air and lowering the dew point. When it is hotter outside, the water will evaporate from the ground faster and the dew point raises. And when it is hottest outside, there is more wind and the mixing of low altitude and high altitude air lowers the dew point. Dew point is dependent on temperature. Dew point is a more reliable indicator of humidity because it is not changed by a change in air temperature and doesn’t fluctuate much throughout the day.

Wind Speed :

Wind, which is as called as air movement, is integral to all types of weather conditions. Wind speed is caused by air moving from high pressure to low pressure, usually due to changes in temperature. There are likewise connections to be found between wind speed and wind course, strikingly with the weight inclination and landscape conditions.

Wind Direction :

Wind direction is obtained by the [direction](https://en.wikipedia.org/wiki/Direction_(geometry)) from which it initiates. For example a wind coming from the south is shown as 180 degrees and the one from the east is 90 degrees. Air pressure, which is largely caused by differential heating of the air by the sun and ground conditions, controls the way air flows.

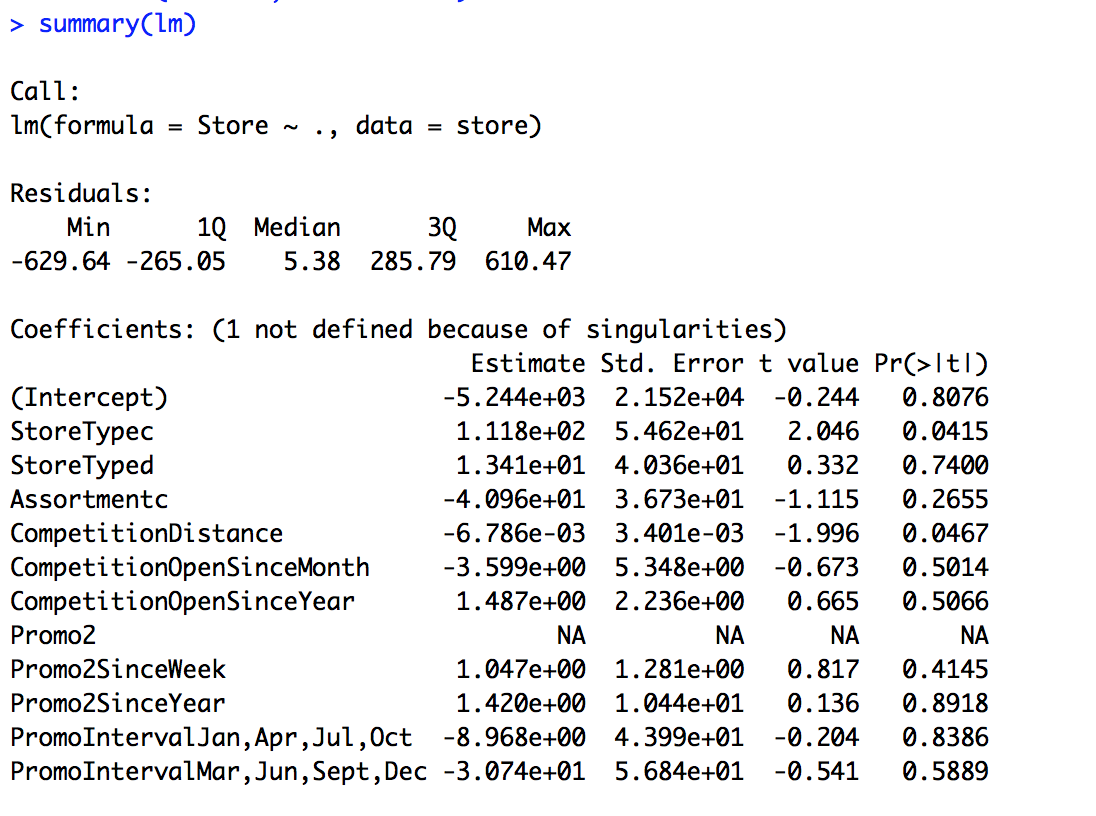
Visibility :

Visibility is a measure of the distance at which an object or light can be clearly seen . visibility affects all forms of traffic, aviation and sailing. If the temperature is warm, the atmosphere has a greater capacity to hold water in its vaporous state than if it is cold. Vapour can change into visible water in one other way.

Timestamp:

Timestamp records the time at which the values are recorded. It is of the form mm/dd/yyyy and it also provides the time at which the values are noted.

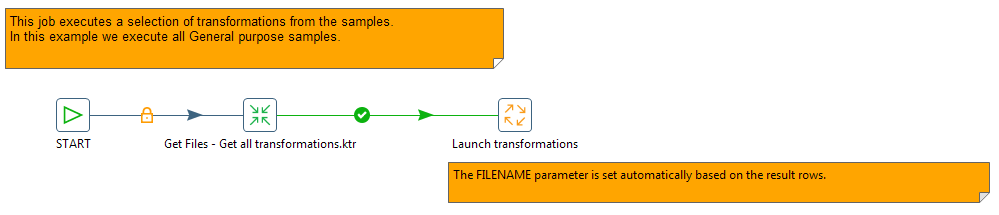
**Method:**

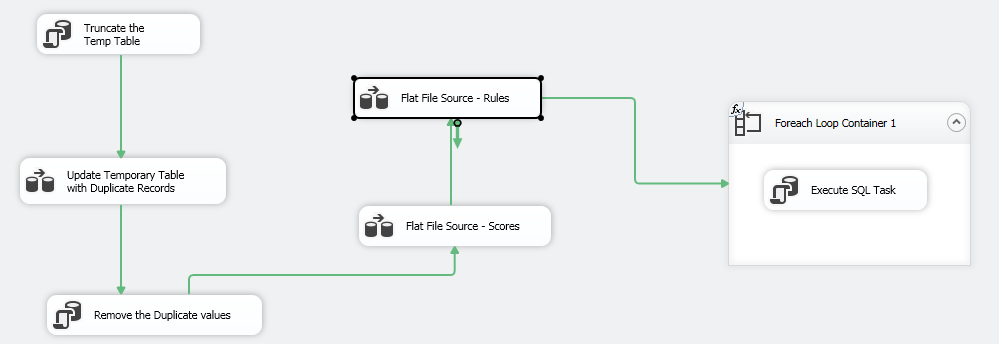
Random Forest model does not overfit. You can keep running the same number of trees as you need. It is quick. Radom it is developed based on a lot of decision tree combinations combined and is also developed by random selection of variables. ****

Let us consider 30 students in a class irrespective of their gender and there are 3 different variables / parameters Gender, Class and height. Out of all the students only 15 play football and now I would want to predict which student would play football as a hobby? In this problem we divide the students based on the sport that they play as hobby from the input. This is where we use decision tree model, It will divide the students based on all the variables in our case there are 3 variables defined.

This creates a homogeneous group of students based on the inputs provided, Gender can be a good variable to identify the best set of or group of students who is likely to play football.

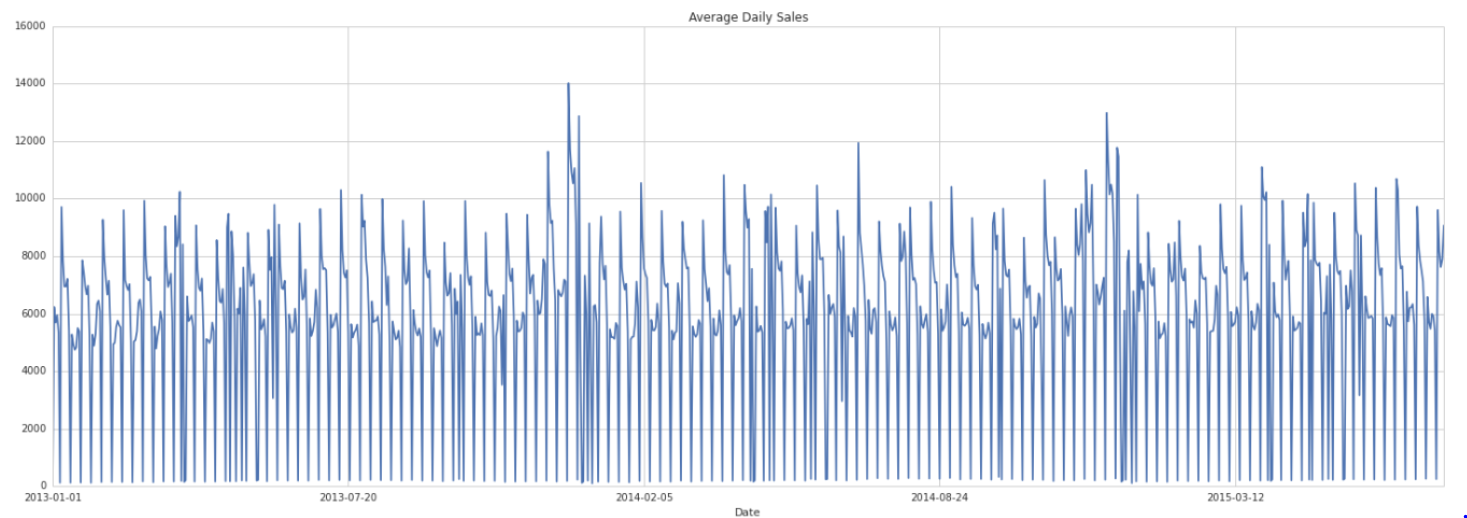
The method uses high end ETL tools to maximize the efficiency of the data flow and minimize the data lose, In my case I have use Pentaho as my ETL tool to extract the data from the csv file for the stores data and this set is derived from a kaggle competition to identify the food sale over a period of time to identify the store with the maximum sales and minimum customers.



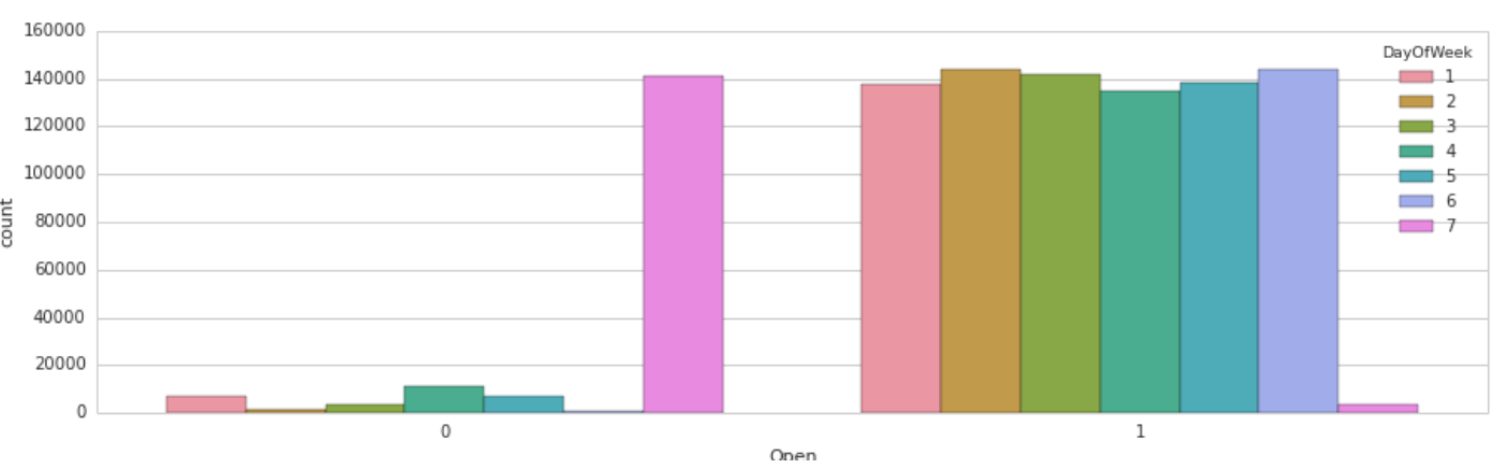
The SSiS version of the ETL load is gived below which gives us a clear idea of the package and shows us the useof each job step and in each job step there are multiple transformation step. The transformation is the phase where the data manupilation happens to the data and this can of any form for example if a new row is to be inserted in to the dataset or change the data type of the data or even calculation of the data which is called as the calculated columns.

**Results:**

In my case study, I have used the weather data to identify the set of people who are likely to purchase a food item on a given day. The variation provided in the model will identify the odd day on which there are a certain set of people who are likely to buy the product this variable are properly measured to indicate that it can be a promotion sales on that day.



This gives us a possibility to analyze the mood swings of the customer and we will also be able to track the seasonal distribution of the sales and the impact of seasonality on various occasions. The graphs give us a real picture of the various factor influencing the change in the customer mood and if we can track the sentiment analysis from twitter data we will be able to understand the emotional situation of the customer and the factors such as online reviews, comments and type of emotions running through the customers on a given period of time. The weather can always be an influencing factor when it comes to mood swings, From the data we can also find the variance with respect to day and night time for the customers to sales ratio to understand when are the customers likely to purchase large amount of products.



The above graph suggests the count of the customers coming on a weekday compared to a weekend and when the store is closed, from this example we are able to understand that there are a lot of customer not able to purchase on Sunday mainly because of the early close and it says that the customer missed is likely to be more than the customers on the weekdays.

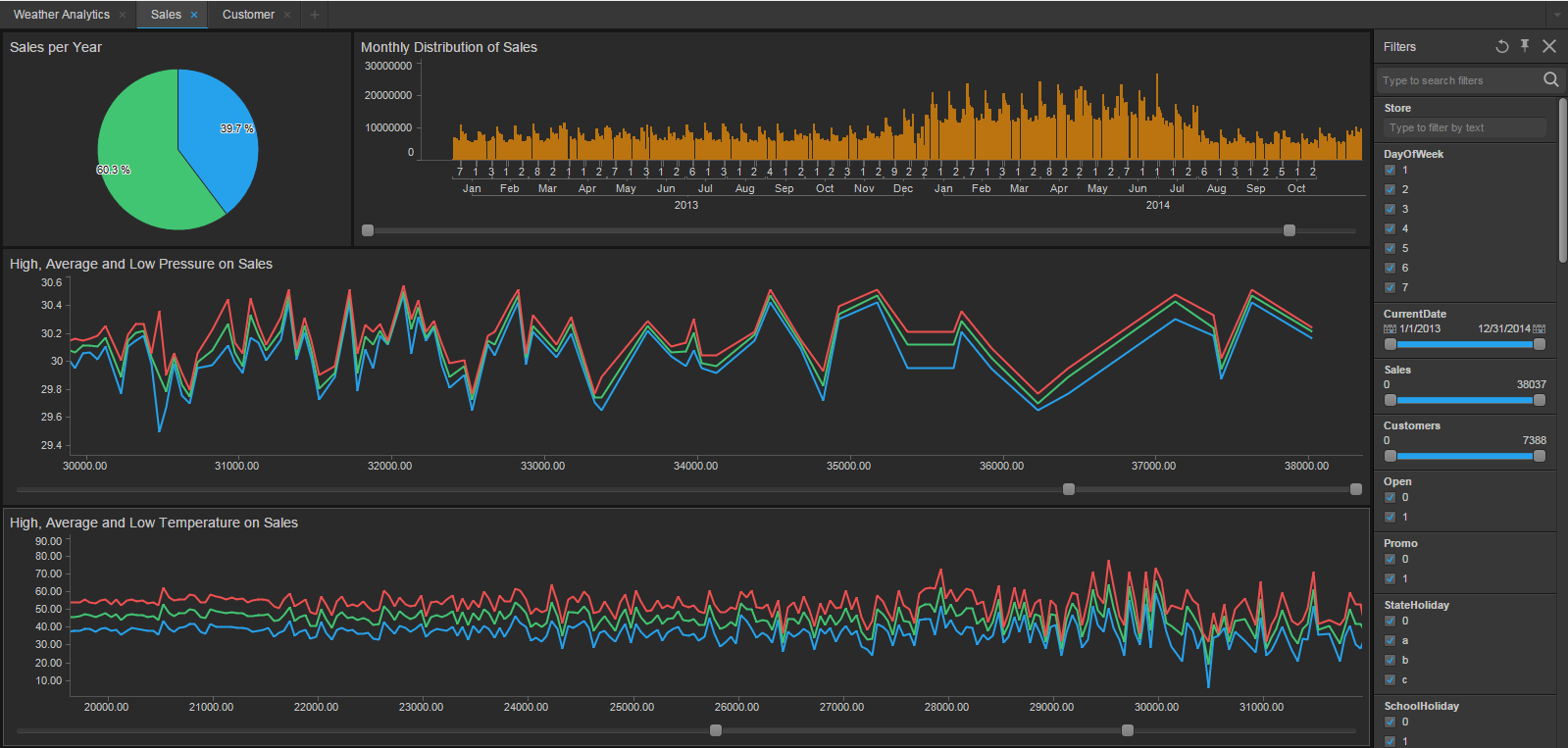
I have implemented Spotfire technology as a medium to interact with the users, Spotfire does a good job to communicate with the end user and drill down even further if there is a need to identify the record/data for the analysis. In my case the data from the below screen-print shows that the sales is plotted for each year and the month icon on the x-axis can be converted to years or dates by moving the scroll bars to the right and left respectively this will aggregate the data and show a overall result of the dataset. Then there is marking which will help to map across different dashboard or graphs to deep further into the dashboard, In the above case the if we select any specific month or a year the below below “Customer per sales” changes according to the selected month. In other words it will show the sales over customer over that period of time. The selection of that range will help us to further narrow down the sales in the promo code pie chart.

It identifies if the customers use promo code during the purchase and helps us to identify if the promo code are helping the stores to get more customers.



Similar to the graph above, the below graph show us the power of spotfire to business needs, The filters on the right will help us limit the data flow in to the visuals, If for example we select the year as 2013 the the data will be subjected to just that piece of the year, by limiting the graph to a certain length of time or a specific store number we will be able to limit the data and show the graph for just that particular piece of the company or the store, The graph below explain the behavior of Sales with respect to temperature and pressure.

The pie chart at the left top is the year which gives us the ability to compare between years and also help us narrow the data to just that particular year. If we select any specific year the the top right graph “Monthly sales distribution” also changes according to the selected data. By selecting the month or the days we can narrow the results to that particular piece of data and see how the temperature has affected the sales of the company/store.



**Comparison with my competitor:**

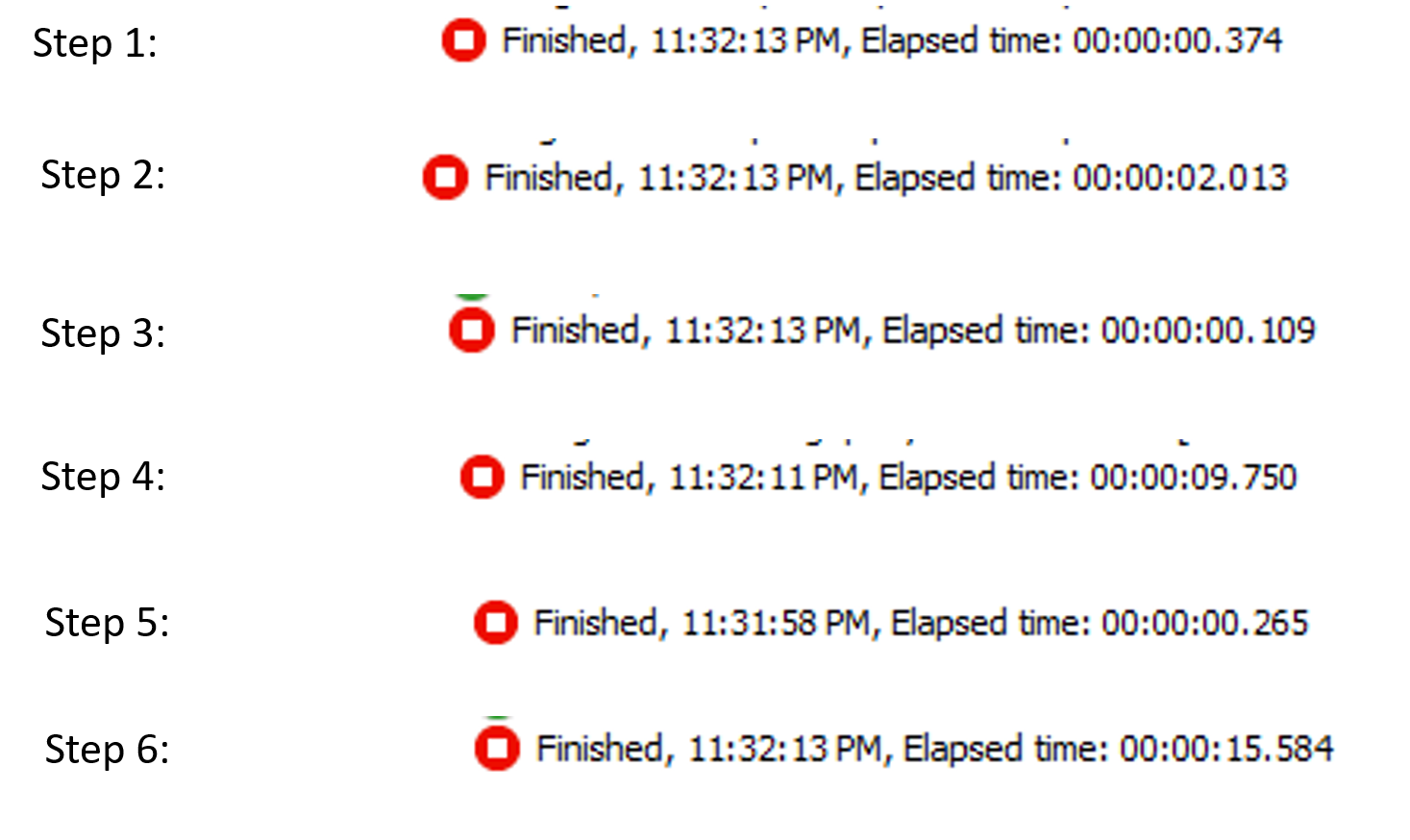
Competitor:

* Slower because of the older technology and no ETL tools used
* Simple Graphs from QGIS
* Hardly to work with live datasets and real time twitter data
* Too much complications on the methodology

My Work:

* SSIS and Pentaho are used as an ETL tool for performance
* Results are in terms of interactive Dashboards
* Easier to perform sentiment analysis
* Simple algorithms matches the same accuracy

**Performance:**



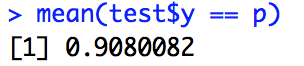
The above screen shot tells the performance of each step and the red icon shows when the process is last stopped and the number indicates the duration of the transformation.

**Conclusion:**

When a product comes to market the demand and supply of the product determines the price of the product, this can be determined in two types; Processed and un-processed food. Processed food are something which isn’t grown directly or it is a derived product of the un-processed food. Un-processed food are grown from the field and should be directly taken in the raw form.

Predicted result:





The test case is divided in to multiple pieces by randomly selecting the subset by 10% of the data and run the regression on each case to identify the accuracy and the highest is 90% and the lowest was 60% and calculating the average from the results, we had a mean average of 75.6%. The total accuracy by the random forest model from my result is almost 76%, which is nearly as equal as my competitor whose accuracy reached to 78.7%, The difference is mainly because of the dataset which can slightly skewed.

* The Data models used by the competitor can’t be implemented in business world because the structure is not designed for data marts and my process is faster when we are working with data lake due the ETL tools .
* Dashboard are developed on the end product to interact with the results and drill down the data if necessary
* Big data technologies are used.
* Result can be further enhanced by using the real/live data as the dataset will not be skewed and this would enhance the performance and produce better outcomes.