Amarcord Case

Business Analytics 2 - Professor: Roberto Garcia-Castro

TEAM 2

Andreas Frey
Andres Sanchez
Eduardo Fernandez
Emma Lotz
Enrico Fausti
Jose Conde
Veronica Pereira
Yuyan Lai

IESE Business School - MiM Class 2020

EXECUTIVE SUMMARY

Introduction

Combatting money laundering has been a growing concern for financial institutions and governments. Some techniques to detect very high transaction amounts can be implemented relatively easily. However, when the criminals of money laundering use techniques that are less obvious, it becomes much more challenging for institutions to notice and detect suspicious frauds.

Method

In the case of Amarcord, after cleaning and imputing missing data using the trend and seasonality analysis, we have been able to build a forecasting model using the Holt Winters analysis. Thanks to this model, we were able to predict out-of-sample data and create a 90% confidence interval for all data from 2007 to 2013. The model has a R-square parameter of 0.72.

Results

The results of our predictions helped us identify 5 out of the 84 months studied whose total amount of incoming wire transfers that are above the 90% confidence interval. These are the incoming wire transfers we would report as a Suspicious Activity Report (SAR) to the U.S. Financial Crimes Enforcement Network (FINCEN). All these suspicious data have amounts below the $500,000 threshold usually used at Amarcord to detect fraudulent transactions.

Implications

Therefore, we believe the management of Amarcord should review their anti-money laundering strategy using predictive analytics, since it gives more reliable and complete results.

REPORT

1. DATA CLEANING

For the data sheet *November 2010 Wires*, we decided to clean the data starting with deleting the outgoing transactions (OUT). Indeed, we are interested in the incoming wire transactions since it is the type of data we have for all the other months from 2007 to 2012. This way, we eliminated 10,670 transactions of our dataset. Then, we decided to eliminate all the cancelled transactions with transaction IDs containing "$C", since we do not have further information about the reason why they were cancelled by the bank (650 transactions). It could be because of a system bug or any other reason that is not given. Furthermore, we decided to delete the extreme outliers and the transactions containing text instead of numbers in the amount.

By summing up the remained incoming wire transfers values, we obtained the value of total incoming wire transfers for November 2010 of $396,812.2.

For the data sheet *Wires By Month,* similarly, we decided to remove the data that are non-numeric and obvious outliers. In addition to these, for the month of August 2010, we subtracted the amount of the proceeds from the sale of the warehouse, since this $37,900 income is an exceptional transaction; This way, we found that the total incoming wire transfers for the month of August 2010 is equal to $456,300. For the month of December 2010, we deleted it because this small amount of income is affected by the fact that a major store of the client was flooded and closed for business. We consider these two amounts not very representative of actual business performance, and as we will use the wires by month data to create a predictive model, we believed it more relevant to subtract or remove these amounts.

2. DATA IMPUTATION

In order to impute the missing data, we decided to use a regression imputation strategy. Before implementing the filling, we observed that the data set reflects some important features. First, they are clearly time series data. Second, within any given year, the collected data exhibit a consistent fluctuation pattern, with the transaction amount in some months generally higher than that of the others. Third, when comparing the data on a year-over-year basis, we noticed that there is hardly an obvious evolvement. The numbers are not necessarily in proportion to periods. Therefore, we decided to apply the classic Trend and Seasonality approach, particularly the additive method, to handle such time series data.

Following the steps of Trend and Seasonality analysis, we firstly created a new variable named "Series" by filling in the sequential number from 1 onwards. Then, we ran a linear regression between the dependent variable (Y) Transaction amount and Series (X). We obtained the Trend values by using the prediction function. Afterwards, we calculated the additive deviations, and the monthly mean of them with the help of group by function. These mean numbers are the seasonal factors in our model. Then, we used the obtained Trend and Seasonal Factor to forecast the transactions for the given periods, with the additive formula Predict = Trend + Seasonal Factor.

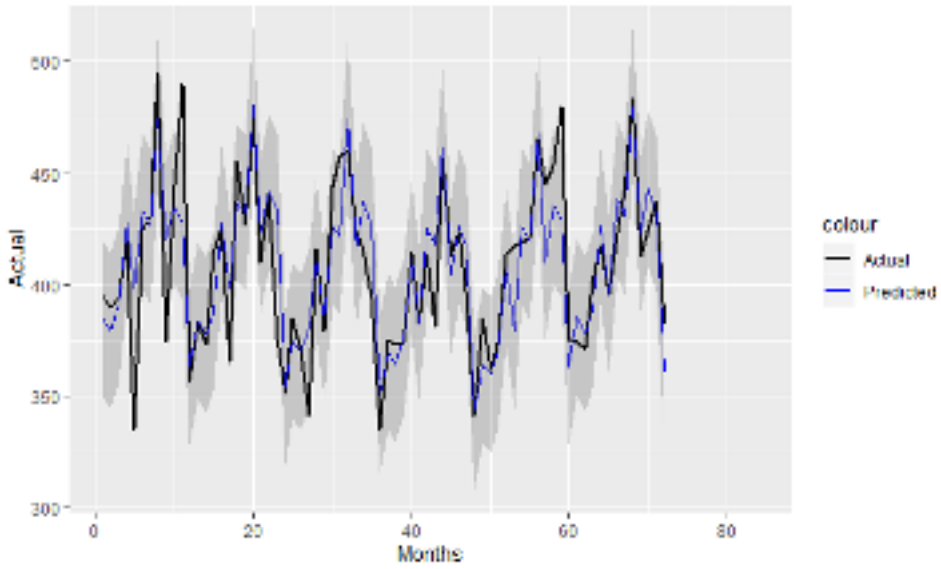
Finally, we filled the missing rows, namely those we deleted in the data cleaning step, by using the predictions obtained from our Trend and Seasonality analysis.

3. BUILDING MODEL

After cleaning the data set, namely using predictions obtained from the aforesaid Trend and Seasonality analysis to fill in the missing data and replace the typos and outliers, we started to build the model by adopting Holt-Winters method. Our rationale of choosing this model rather than others is that Holt-Winters analysis is able to learn and replicate an observed pattern in a continuous manner, whereas other models like ARIMA tend to have a static algorithm from the moment the model is built. Such continuously adapting nature is beneficial in this case because it allows us to better cope with the data which have innate complexity of trend as well and a changing nature.

To start, we transformed the clean data set at hand into time series that can be deployed in Holt-Winters analysis. We also separated the data set into two subsets, training and testing. Afterwards, we ran the Holt-Winters analysis with the training data. Just as we observed and decided in the Trend and Seasonality analysis, here we chose the Additive method over the Multiplicative method.

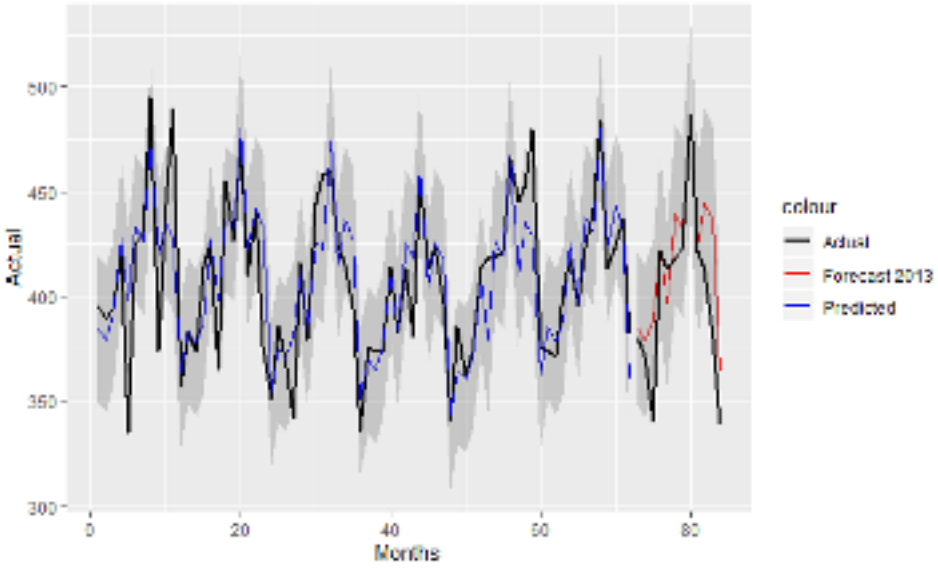
With this model, we obtained the predictions for transaction amount from 2007 to 2012. After obtaining the predictions, we compared them with the actual numbers to see how these two sets of data are deviating from each other. To have a direct visual view, please refer to the below plot.



We also computed the R square which is 0.69. The R square parameter suggests that our model explained 69% of the variation of the independent variable.

4. FORECASTING/TESTING

Another way to test how reliable is our model is to apply it to a new set of testing data. To this end, we forecasted the transaction amount in the year 2013 by using the Holt-Winters model generated from the training data. We then plotted the predictions together with the actual amount, as shown in the below graph.



The R square parameter is 0.72, meaning our model explained 72% of the variation in 2013. As we can see, this R square is higher than the one obtained on the training data. This could be explained by the larger number of outliers (potential months with incoming transactions fraud) that the training data has compared to the data that we are testing. The RMSE parameter is 62.68, meaning on average, our predictions deviate from the actual transactions by 62.68. We believe that the RMSE in absolute terms is large; but given the number of outliers and that we are working with a very small data set, it is acceptable.

In the table below we compared the results from the three methods that we used for the forecast. Arima is the method that performed the worst from the three while the performance of Trend Seasonal is arguably better than the one we obtained from our Holt-Winters as it has a lower RMSE. At the end we decided to use Holt-Winters as it has a larger R square. Also, we believe that with more ongoing data input, Holt-Winters will be able to yield a better result.

Method	R_squared	RMSE
Trend Seasonal	0.7156595	56.82068
Arima	0.7038627	57.04532
Holt winters	0.727955	62.68024

5. ANALYSING

Using a 90% confidence interval to detect anomalies, we found that 12 out of the 84 months used in the analysis have a total amount of incoming wire transfers that lie outside the 90% confidence interval. These months correspond to months number 5, 9, 11, 23, 27, 31, 43, 53, 57, 59, 75, 83, as seen below.

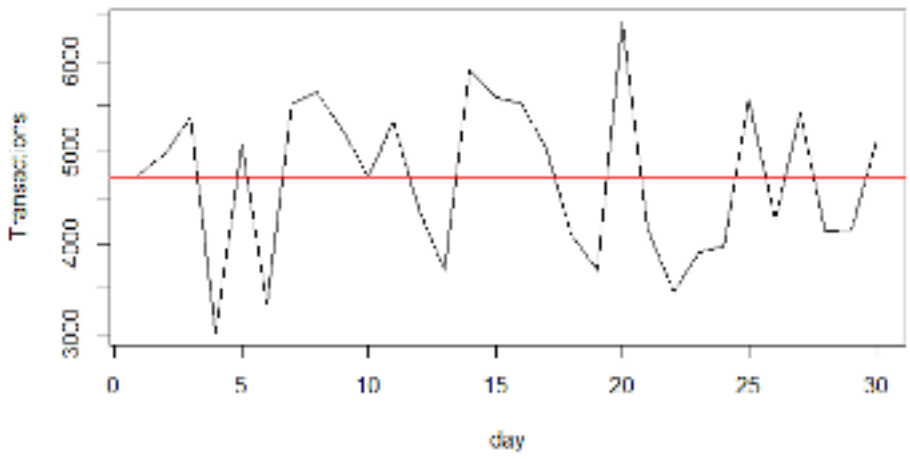
	Actual	Predicted	Upper	Lower	Inside	index
5	335.25	397.0291	432.1205	361.9377	Out	5
9	374.27	418.9631	454.0545	383.8717	Out	9
11	489.50	428.5039	463.5953	393.4125	Out	11
23	375.75	432.8049	467.8963	397.7135	Out	23
27	341.44	379.8993	414.9907	344.8079	Out	27
31	457.50	421.6903	456.7817	386.5990	Out	31
43	381.00	417.7865	452.8779	382.6951	Out	43
53	417.70	379.6021	414.6935	344.5107	Out	53
57	444.91	409.3679	444.4593	374.2765	Out	57
59	479.80	428.3016	463.3930	393.2103	Out	59
75	340.40	386.9006	424.3279	349.4733	Out	75
83	379.80	435.1035	478.8024	391.4045	Out	83

We would report only the amounts that are above the confidence interval as Suspicious Activity Report (SAR) to the U.S. Financial Crimes Enforcement Network (FINCEN). Therefore, we would report the amounts of months 11, 31, 53, 57, 59, which corresponds to periods November 2007 ($489,500 total incoming wire transfers), July 2009 ($457,500), May 2011 ($417,700), September 2011 ($444,910) and November 2011 ($479,800) respectively.

Looking at November 2010, the total amount of incoming wire transfers lies within the 90% confidence interval. However, another assumption we could have made is that the bank cancels incoming wire transfers when they have a fraudulent component. Therefore, if we do not take out these cancelled transactions ($C) of the dataset, the actual amount of incoming wire transactions for November 2010 would be higher by $38,040.11 so the total amount would be $434,852.3 which still lies in the confidence interval [385.43;455.6] so we would not report it as SAR.

Additional findings

As the data from November was interesting, we tried to break it down to find potential transactions that might be fraudulent. To find out which transactions might be fraudulent we decided to use a score system to weight the different results using chi-sq scores ((transaction-mean of transactions)^2/ variance) to assign a score to each one of the transactions. Using the function score and setting the confidence limit to 90% we found out that 6% of the scores laid out of the 90th percentile which we could consider as extraordinary transactions. As the data was interesting, we plot the number of extraordinary transactions per day.



As there was a big peak on day 20 we isolated that data to observe how the extraordinary transactions of that were and we found out that a lot of the larger transactions were done very close to each other and in a lot of cases large quantities of Incoming and Outcoming transactions happened in the same minute or minutes away from each other meaning. Looking at these indicators I would guess that fraudulent transactions (extraordinary scores in our analysis) are highly correlated with fraudulent transactions and that the closeness of large transactions are highly correlated with fraud.

Limitations and recommendations

The dataset Wires By Month available for analysis is relatively small and we would get better results with a larger dataset.

Moreover, there are a lot of outliers in the dataset and a lot of missing data which has hampered the accuracy of our models.

For the future, we would advise Amarcord to keep a better track of its transactions and to implement new policies in their fraud-detecting methods using predictive analytics like presented in this report since it gives more reliable and complete results.