# K-Means Clustering Approach for Improving Financial Forecasts

Țole Alexandru - Adrian
*The Romanian - American University*
alexandru.tole@gmail.com

## Abstract

*The following paper treats both types of forecasting: qualitative and quantitative. It highlights the importance of using both of them in order to achieve more accurate forecasts.*

*It shows the flaws of quantitative forecasting when applying simple regression on large sets of data. Also, by using advanced data analysis techniques, such as Big Data algorithms, the results of the quantitative forecasting can be drastically improved and it can be worthy of taking into consideration when drawing the conclusions.*

*K-means algorithm it proves to be very effective when a quantitative forecast needs to be done. By using it we can successfully execute "drill-down forecasting" into specific activities.*

**Key words:** clustering, k-means, quantitative, qualitative, forecasting
**J.E.L. classification:** G17

## 1. Introduction

A key component in every company development is represented by investments. These can be done with internal and external financial resources as well. Ideally investments should be done by using the company's financial resources and not applying for bank loans, for example, that can generate other costs which can put a lot of pressure on the organizations budget. In order to be able to sustain any investment during its implementation by using internal financial resources, forecasting the revenues of the company for that period is crucial.

## 2. Theoretical background

The forecasting can be: *judgement (qualitative) forecasting* and *quantitative forecasting*. The *judgement forecasting* is based more on intuition and the experience that you have as the owner/CEO of the company. Is often used when quick decisions are to be made and it has a higher rate of failure if they are not done by experts. Also, this type of forecasting is used when historical data is not available or it cannot help the type of decision that needs to be taken. It is used for medium and long-term decisions by consulting focus groups, expert opinions or even doing historical analogy in order to determine trends.

On the other hand, the *quantitative forecasting* is a scientific approach and the results are based on applying techniques and formulas on historical data. By using this we can achieve better results and predictions even if we are not experts in that type of business. This type of forecasting can also be achieved with the help of advanced data analysis techniques that come from different fields and can be adjusted to obtain better results. For this type of forecasting the quality of data is very important because it has a direct impact on the accuracy of the result.

To be able to have an accurate forecasting is indicated to combine both types, because where one lacks of information the other can fill the gap. Even if the company is newly created, the historical data can be obtained from companies that have the same profile in order to achieve also a quantitative approach. In this case, the lack of experience might also be present and for that we can choose to ask experts for their opinion in order to achieve judgement forecasting also. In this way, we can make a forecast based on qualitative and quantitative information that can give us a better overview of the situation.

In any case, a sound financial forecasting is the backbone of every company direction. Based on this, management can plan and make strategies in order to correct a bad course and also to achieve the goals that has been established.

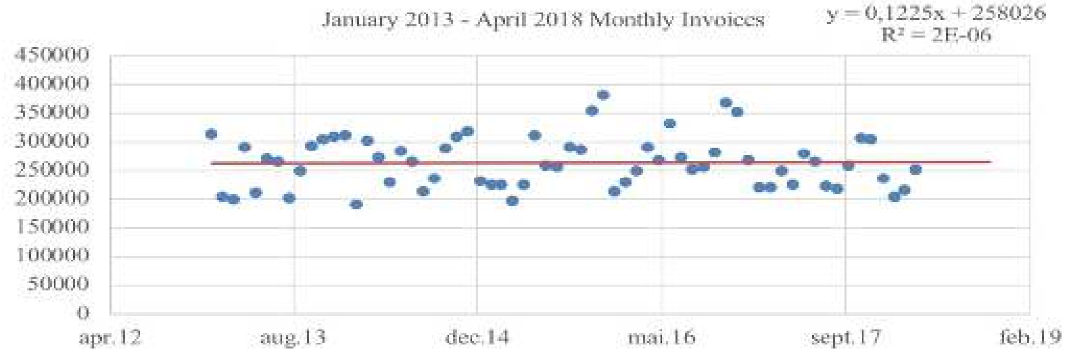
## 3. Quantitative forecasting using simple regression

By using this type of quantitative forecasting can be very helpful if there is a single casual variable that can be controlled. The form of the simple regression is "y = a + bx, where "y" is the variable to be forecast (dependent variable), "a" is the constant, "b" is the effect size, and "x" is the causal variable" (Armstrong *et al*, 2018, p.12).

A very important statistical measure when applying simple regressions is R-squared. This represents the coefficient of determination for the regression. Simply put, is the "variance of the predicted values divided by the variance of the data" (Gelman *et al*, 2017).

The R-squared value can be between 0% and 100%. If the value is closer to 100% it means that the regression model is very close to the observation. As an example, I used data from a company that sells services and has plenty of clients that are loyal and come each year and buy the same services with little variation in contract value.

*Figure no. 1. Simple Linear Regression – Forecasting example based on more than 5 years of data (monthly)*

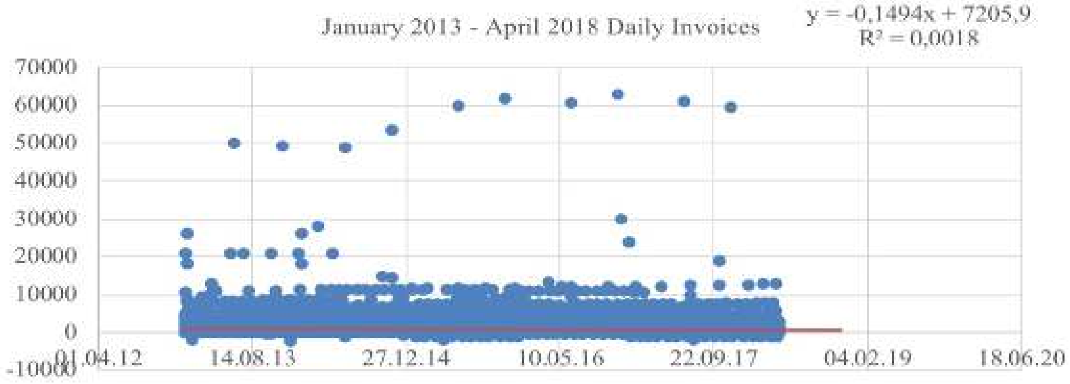


*Source:* Romanian National News Agency AGERPRES: January 2013 – April 2018 invoices

The R-squared in *Figure 1* is so low that it can't be accurately represented. In this case, the regression model is very far from the observation and, in theory, the forecasting that has been done is not accurate at all. One solution, proposed by some publications, for improving this is to increase the quantity of data. For the chart above were used only 64 lines of data, representing the total value of invoices emitted each month from January 2013 until April 2018.

Applying the solution proposed, in *Figure 2* there are more than 19.000 lines of data that represents the invoices for the same period as *Figure 1* and the outcome is the following:

*Figure no. 2. Simple Linear Regression – Forecasting example based on more than 5 years of data (detailed)*



*Source:* Romanian National News Agency AGERPRES: January 2013 – April 2018 invoices

In this case the R-squared has slightly improved but is still very close to 0% which means that drawing conclusions after the forecast is very hard if we are relying only on quantitative forecasting. Even if we apply logarithmic or polynomial regression the R-squared is not improving by much and the forecast is still unclear.

## 4. Qualitative analysis

In order to obtain a better forecast is necessary to include qualitative data at this point. In this case is necessary to obtain relevant information about the company profile, clients profiles, types of services, recurrent invoices etc. The best way to do this is by talking to the persons directly involved.

*Table no. 1. Collecting qualitative data*

| Relevant topics to consider | Observations |
|---|---|
| Are there any contracts signed with monthly invoices or just single invoices for one-time service/product? | Both types of invoices are present in the chart, but most of the incomes are from contracts with monthly invoices. |
| How is the overall trend in the industry? | The sales of the core products in the industry are decreasing national and worldwide. |
| What is the profile of the clients? | Clients are both from public and private sector. High value contracts are mostly signed with public institutions. |
| How is the competition? | One of the main competitors has serious financial problems and some clients where won back from him. The rest of the competition is tough in some areas. |
| Are there any legislative initiatives that can impact incomes for the company? | The acquisition legislation that impact the public sector puts pressure on the whole acquisition process, slowing it down. The legislation was changed in 2016 and applied in 2017. |
| Are there any other problems that should be taken into consideration? | As stated before, a very large part of the income is from public institutions. These are highly dependent on the national budget and if they get a lower budget also their capacity of contracting this type of services is negatively affected. Also, late budget approval should be taken into consideration because it can have a very high impact on the incomes for the first quarter. There are products that are strictly dependent on the political events going on. (Mostly elections) |

*Source:* Personal observations

By getting the answers to these questions we can see that there is a high chance for a negative trend in the following period. In this case, if there are any investments that need to be done they should be careful supervised and no risk should be taken.

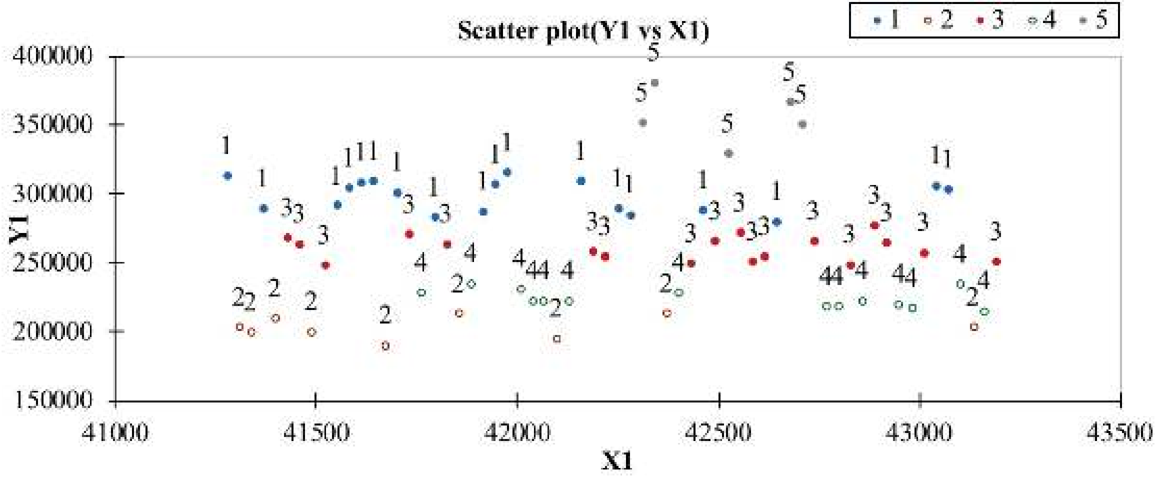
Still, by having all the data to see the bigger picture is not enough. We can conclude that the trend will be very close to what it was until now with a tendency of decreasing but if a bigger investment needs to be done in order for the business to grow the financial forecast must be more accurate.

## 5. Adding K-Means clustering technique to financial forecasting

Clustering is a technique of data mining that groups objects into clusters by following some rules. K-Means is one of the most used algorithms in data mining because it allows us to create groups from very large datasets. The algorithm uses the "concept of Euclidean distance to calculate the centroids of the cluster" (Virmani *et al*, 2015). This means that the data is going to be grouped around the defined centroids.

By applying this algorithm to our data represented in *Figure 1* and dividing it in five clusters we will obtain five groups that are centered around the invoice value – *Figure 3*. Starting from this point we can forecast the possibility of invoicing specific values – *Figure 4*, we can also determine the trends of selling products by correlating the invoice with the product and even more detailed forecasting.

*Figure no. 3. Clustering Invoices – Creating five groups around the invoice values*



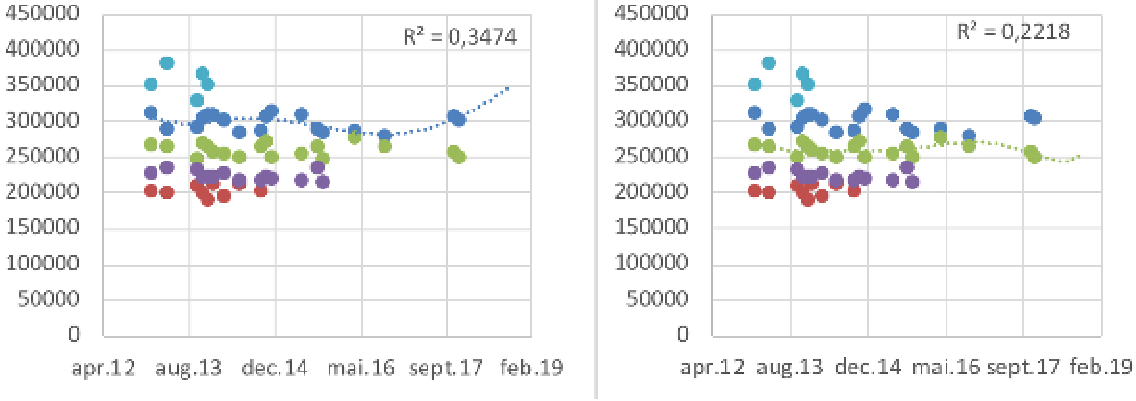
*Source:* Romanian National News Agency AGERPRES: January 2013 – April 2018 invoices

We can see in *Figure 3* how the groups are formed around the values represented in Y1. For those values there were five groups defined starting from the lowest invoice value to the highest. By calculating the Euclidean distance we can determine the centroid of each group and we can identify its members.

Doing so, we can start forecasting on invoices values and we can determine the trends of those in order to know on what incomes we can count on when we are doing investments. I took the median values – *Figure 4* - of the five groups and applied polynomial regression in order to see the trends. For both forecasts the R-squared value is much better than *Figure 1* and we can now say that we can also rely on quantitative data when drawing the final conclusions.

By analyzing the results from both quantitative and qualitative data we can now assume that the trend of invoicing median values is growing but we still need to be cautious because there overall qualitative data is forecasting a negative trend.

*Figure no. 4. Forecasting Invoice Values – Forecasting on two out of five groups of invoices*



*Source:* Romanian National News Agency AGERPRES: January 2013 – April 2018 invoices

## 6. Conclusions

Predicting the company’s trends will always be of great importance for managers. This can only be achieved if the approach is correct and is done by experts.

Financial forecasting will always be challenging for any expert even if is enough information to draw conclusions from. The best way to predict what will happen with the company incomes is to use both quantitative and qualitative data. By doing so, the margin of error will drastically reduce and managers will be more confident with their decisions. As we saw earlier, quantitative data, no matter how much we have or how qualitative it is, is not enough to create pertinent forecasts. There should always be qualitative forecasting also be taken into consideration for better understanding the quantitative forecasting results.

The quantitative forecasting results can be improved by using techniques from other disciplines like Big Data algorithms. Even if K-means is a very simple algorithm is also very efficient and can create great added value to any type of forecasting. In my simulations this algorithm is proved to be limited only to "drill-down forecasting" which is useful for better understanding certain trends on specific activities.

## 7. References

- Armstrong, J.S., Green, K.C., 2018. *Forecasting Methods and Principles: Evidence-Based Checklists*, [online] Available at: < https://goo.gl/k28NVf > [Accessed 04 May 2018].
- Gelman, A., Goodrich, B., Gabry, J., Ali, I. 2017. *R-squared for Bayesian Regression Models*, [online] Available at: < https:// goo.gl/cpFnVN > [Accessed 05 May 2018];
- Virmani, D., Taneja, S., Malhotra, G., 2015. *Normalization based K Means Clustering Algorithm*, [online] Available at: < https:// goo.gl/7ovLFR > [Accessed 06 May 2018];