Università degli studi di
Milano - Bicocca

Decision Models

Final Project

Markdowns in a supermarket chain: a decision study

Authors:
Alex Ceccotti - 790497 - a.ceccotti@campus.unimib.it
Davide Pecchia - 793290 - d.pecchia1@campus.unimib.it
Michela Sessi - 777760 - m.sessi@campus.unimib.it

June 17, 2018

Abstract

Applying or not markdowns is an important choice that a supermarket necessary has to make in order to keep its loyal customers and acquire new ones. In this project, the aim is to help the market's management to decide whether applying markdowns in a certain week is a good or a bad choice. Using data regarding an American supermarket chain, after a first manipulation performed to create the dataset for the analysis, a decision tree has been built using the most significant variables. Some of these (in particular the variables *Weekly Sales* and *Ratio* needed to be predicted; it was done by implementing a predictive model for each one of the 2 variables. Also the variables *CPI* and *Unemployment* rate had to be estimated. For this reason, an ARIMA forecasting model was used. Once the final tree has been implemented, the model has been tested on a sample of data in order to verify the performances. The quality of the data used was not optimal: the model could be improved with few changes in the data to enrich the information to use during the building process. Anyway this study is an optimal starting point to help stores in the decision process about the markdowns application.

1 Introduction

The main purpose of a large supermarket is to sell and gain as much as possible. For this reason marketing has developed ideas that lead the customer on some occasions to be more attracted to the purchase. We are now all used to seasonal sales, flyers advertising with offers, to 3x2 or to shop earning points and making the purchase a game with prizes. But when should the market offer discounts to attract customers and when the discount becomes a loss of earnings? If the customer buys only products on sale the choice of marketing does not work. It is not enough to attract more customers. Otherwise the point is to increase the revenue. Depending on various factors, offers can generate a different attractiveness: being in conjunction with a week of holidays, the weather and unemployment rate are just some of the variables that could affect the income of the supermarket.

1.1 Goal of the project

The purpose of this study is to optimize the profit depending on the store of an American supermarket chain. In particular, a decision model has been developed in order to understand if it's convenient to apply markdowns or not for the store, which would still earn from its usual customers without special offers. The choice is diversified depending on the week of the year and the size of the store; several other external variables are taken into account.

2 Datasets

For the analysis has been used three datasets (.csv files) available on the Kaggle platform [1]. Data are related to 45 stores during 143 weeks of an American supermarket chain.

- The *Features* dataset contains information about the store number, date, temperature, fuel price, 5 different markdowns, the CPI (Consumer Price Index), the unemployment rate and a variable called *IsHoliday* (a binary TRUE/FALSE variable which indicates if the date corresponds to an holiday period or not).
- The *Sales* dataset contains information about the store number, department number, date, amount of weekly sales and, again, the *IsHoliday* variable.
- The *Stores* dataset contains information about the single store. In particular, for each store number, it's possible to find out the type and the size of the store.

Datasets are related to the sales from 05/02/2010 to 01/02/2013. Unfortunately the *Features* dataset does not contain all informations: data are available only from 11/11/2011 to 26/10/2012.

2.1 Data Manipulation Processing

In order to prepare the data and create a single full dataset to be used within the analysis, after importing all the data in R, an aggregation has been made on the *Sales* dataset. In fact it was more useful to gather the sum of the weekly sales for each store and not divided for departments, because

it was an information not available for the other variables. Subsequently, all the 3 datasets has been merged to create the final one, collecting all the information for the analysis. From this merge data were available from 11/11/2011 to 26/10/2012. To enrich it, functionally to build the model, some transformations have been done and some new variables have been created:

- *Date*. In order to manipulate data, the date has been transformed from string to american date format: %Y%m%d
- *AfterHoliday* and *BeforeHoliday*. Two binary variables like "IsHoliday" have been created to define if a certain week was before or after an holiday period.
- *Season*. A variable that identify the season was useful. In addition that feature has been binarized in four dummy variables one for each season: Autumn, Winter, Spring, and Summer.
- *WeeklySales*. Obviously sales depends on size: the more the store is big, the more the sales increase. In order to get an objective feature that does not depends on size of the store, a new variable was necessary: *Weekly Sales* has become proportional to size. For our model it was useful to treat gross sales, so weekly sales have been added to total markdowns (sum of reductions in price).
- *Ratio*. Ratio was equal to the sum of the markdowns divided by the not transformed weekly sales in addition to the sum of the markdowns. In this way it is possible to have an opinion on the impact that the offers have had on sales by week and by store.

It has been obtained a unique dataset with 2295 rows (45 stores x 51 weeks) and 19 variables.

3 The Methodological Approach

It was decided to use a decision tree to choose whether apply or not the markdowns. The decision is "apply all markdowns" or "don't apply any markdown". It was not possible to conclude anything about the single markdowns because no specific information was released about the impact they

have on the sales if taken alone. The decision tree should take in input the *Weekly Sales* and the *Ratio* explained in the section above and should give in output the decision to make. Anyway, at the beginning of the week those informations are not available, but they are estimable. For this reason it has been chosen to develop two regression trees in such a way that, taking some variables in input, they give in output the estimate for the *Weekly Sales* and the *Ratio* between markdowns and sales. The next step was to probabilize the week's average temperature. This choice complicates the problem because in the decision tree there will be more branches with a given probability. On the other hand, to make the problem not too twisted, the other variables were assumed to be deterministic. Finally, it was hypotized that, if the markdowns were not applied, the store would lose a percentage from the estimated *Weekly Sales*. This percentage is calculated taking into account the *Ratio* and the demand's elasticity. In this way a trade off between apply markdowns (and so making lower prices) and not apply markdowns (and so losing some clients) has been generated.

3.1 Predictive tree for Weekly Sales

It was necessary to predict an evaluation about the gross sales of a certain future week. For that reason it has been performed a tree that has as target *Weekly Sales*. To choose the input variables for the regression tree and to set a reasonable cp parameter, the model was tuned trough the caret package using a 10 folds cross-validation approach and a grid search. Looking at the Importance of the input variables given in output by the caret's train function, it was observed that *Unemployment* and *Cpi* are the most important ones. The third one (*Temperature*) had as value $\frac{1}{3}$ of the *Unemployment*'s importance, so it has been dropped such as the others. The cp parameter was set to 0.01 to avoid overfitting and at the same time obtaining a good RMSE. The predictive tree is shown in Appendix section figure 4.

3.2 Predictive tree for Ratio

It was necessary to build another regression tree in order to predict the values for the variable *Ratio*. *IsHoliday*, *BeforeHolidy*, *AfterHoliday*, *Temperature* and *Winter* were easy to interpret in a regression tree. Temperature above all was also easy to be probabilized in the decision tree. Otherwise, after a tuning step (same approach as the *Weekly Sales* predictive tree), cp was

set to 0.01 to mantain an appropriate relationship between simplicity and fitting. To improve the fitting, the idea was to create a neural network for everyone of the last leaves of the tree. Finding the predicted value in the leaf for every observation was the first step. Subsequently vectors and a list were created and initialized in order to be filled later with the results (the list should have been filled with the 8 neural networks). After that, the neural networks implementing process could start, using only the variables *Type*, *Size*, *Unemployment* and *CPI* to train the model, those ones were the most correlated with the response value. It was used the option *"tuneLenght"* in order to find the best number of neurons and to use the best value for the decay parameter for each neural network. To validate the results a 5-folds cross validation was used (it was a well-chosen number, because some leaves had a numerousness of about 50, so using a 10-folds cross validation was not efficient). So the best neural network was estimated and saved in the list created at the beginning. The R^2 was also calculated to have a measure of the prediction quality. Even if it is not good in this case (about 25%), it is a good improvement compared to basic models. The principal problem was the connection between covariates and response. The predictive tree is shown in Appendix section figure 5.

3.3 ARIMA for *CPI* e *Unemployment* rate

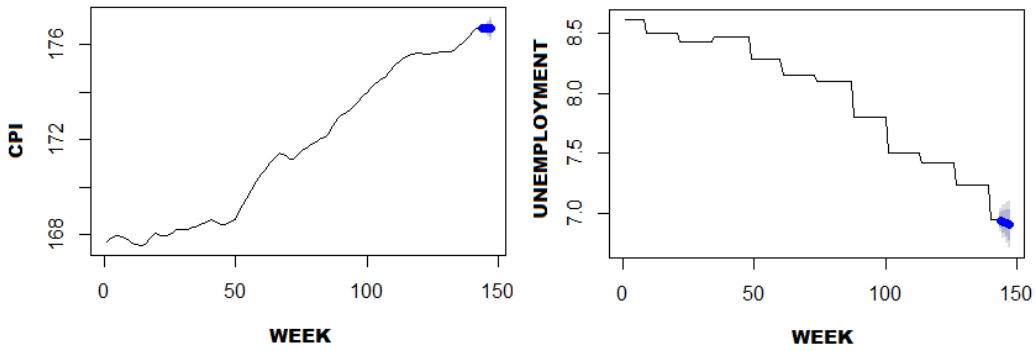

Figure 1: Forecasts for CPI and Unemployment rate

To choose whether apply or not the markdowns it was needed to know, among others, variables like *Temperature*, *CPI* index and *Unemployment*

rate. While temperature is taken as a given value with a certain probability, *CPI* and *Unemployment* must be estimated. For this reason, the best approach is to forecast the future value for these variables trough an ARIMA model [2]. Using the *auto.arima* function from the *forecast* R package an optimal model was automatically found. Considering the *CPI*, an ARIMA(0,2,1) has been obtained, while an ARIMA(0,1,0) with drift was the best one for *Unemployment* (see Figure 1). The forecasts given in output by these models has been used as input variables for the two predictive models explained above.

3.4 Decision tree

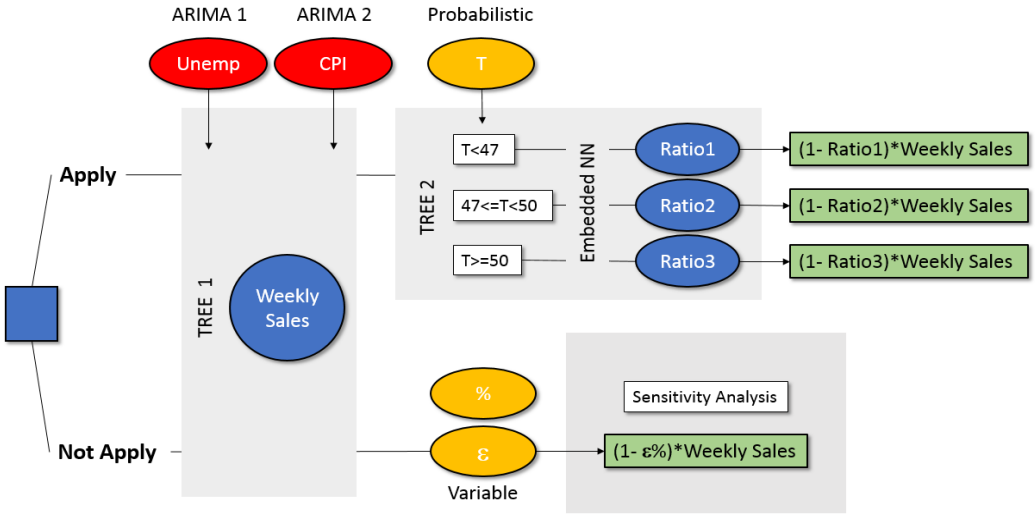

Figure 2: Decision Tree

To evaluate the choice, a decision tree with two branches (Apply and Not Apply) has been created. To make the evaluation, for each branch, the expected value of the store's *Weekly Sales* must be calculated. Obviously the best choice is the one with the highest value. In the first branch, different results were obtained depending on the probabilized *Temperature*. In fact, three different values with a probability associated are given in output: one for $T < 47$, another for $47 \leq T < 50$ and finally one for $T \geq 50$. The probabilities are calculated as a Normal distribution centered in the predicted

Temperature with a standard deviation of all the temperature of the season. Net sales are calculated with the following formula: $(1-Ratio)*WeeklySales$ where *Weekly Sales* are the gross sales given in output by the first regression tree. Otherwise, in the second branch everything is deterministic but some variable are choice of the domain expert. The percentage of discount has been considered: for this case markdowns correspond to 20% reduction, so in order to come back to the full price it has been estimate a 25% of increment (=1-(100/(1-20%))). Another variable has been used: an elasticity parameter ε. The hypothesis, accredited by various studies, is that by not applying markdowns, so increasing the price, there is a loss of customers and therefore a loss of percentage sales; this is estimated by the ε parameter that in the literature changes according to types of products, years and other factors known by the domain expert. In a work of *Journal of Marketing Research* [3] the author studied the elasticity parameter for lots of products and reported the results in a distribution shown in Figure 3. A reasonable ε could be 0.095

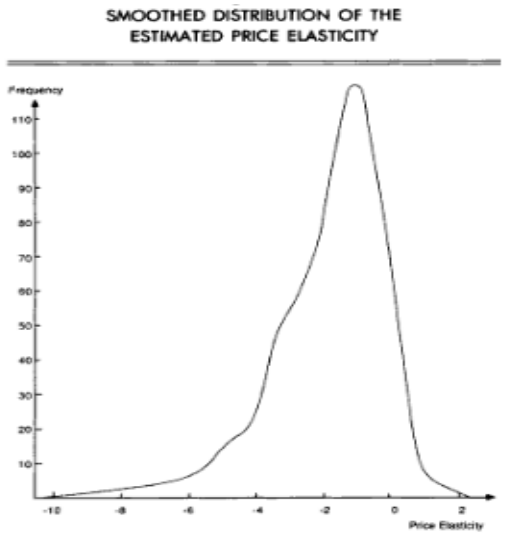

Figure 3: Elasticity curve

[4] if the products are bread and cereals, but this is just an example. The final choice is anyway left to the domain expert. Net sales were calculated with the following formula: $(1-\varepsilon\%)*WeeklySales$. For this work it has been implemented a Sensitivity Analysis with which the expert could decide his approach to the problem, could evaluate the elasticity and could get the

choice to apply or not apply markdowns.

4 Results and Evaluation

In order to test the model, the decision tree has been implemented with variables from the following month of data's study. In particular, for example, 4 weeks has been considered: 2012/11/02, 2012/11/09, 2012/11/16, 2012/11/23. The choice to use this month is lucky because 22^{nd} of November was the Thanksgiving day: for that reason there are all cases for *BeforeHoliday, IsHoliday* and *AfterHoliday*. Characteristics of the store that has been selected are: size = 151315 and type = A. ARIMA models have been easily used for the estimation of indices *CPI* and *Unemployment*. Obviously the season is autumn; to hypothesize the temperatures of this four weeks a random sampling has been used just starting from autumn temperatures. In order to get as output a decisional tree, the elasticity parameter ε is supposed to be equal to 0.095; that value is justified for the food industry that is connected with the supermarket chain problem. In the first week there are not holidays, and *BeforeHoliday* and *AfterHoliday* are False; the second week is before an holiday; the third week is the Thanksgiving one; the last one is after the holiday. With the ε decided, the choice is to apply markdowns at 20% for all the cases except for the Thanksgiving week. Sorted by date the expected value of these net sales are equal to 1239844.76$, 1231837.30$, 1237234.29$. The holiday week gets the choice to not apply markdowns (expected value = 1224166.15$) Anyway a Sensitivity analysis for every week has been plotted for an accurate decision taken by a domain expert, in this way the value of the elasticity parameter could be changed. Decision tree and Sensitivity Analysis of every week are shown in Appendix section from Figure 6.

5 Discussion

From data used to test the study it has been observed that the holiday week does not apply the markdowns. This could be a sensible choice considering the fact that, during the Thanksgiving week, families buy more than in other weeks for the celebrations that await them. What emerged, however, is a lower value for the expected value of net sales. In fact, to estimate the *Weekly*

Sales it has been used a tree that takes into consideration only the *CPI* and *Unemployment* rate variables, estimated by the ARIMA and therefore with very similar values for the whole month; in this way, the most important variable that determine the final result are the probabilized values of the *Ratio*. On the other hand, it would be more effective to have other variables that can more accurately differentiate *Weekly Sales*. The choice works as long as the value of the elasticity parameter is true, if it was set a too large ε the decision will always fall onto the application of the markdowns and this is clearly seen in Figure 10. Furthermore we could obtain different results depending on the percentage of the markdown we are investigating (in this study set at 20%). Remember that the temperature is an average of the week and may not be a reliable representation if the weekly distribution has a strong variance; surely, temperature can not be taken alone into consideration without other atmospheric variables. Finally it has to be clarify that the assumptions are very strong: the work is based on data of the year 2011-2012. Currently, after many years, it is impossible to execute the decision if not obtaining updated data. Furthermore, being the work characterized by a particular supermarket chain, it is difficult to extend it to any market not in the chain without a detailed study.

5.1 Improvements

It would have been better to have different variables such as weather or costumers. Furthermore, having daily details could improve a lot the models performance. With this hypothetical data, it could be also possible to try other models which could be more appropriate for the problem. Moreover, having data about more years would be interesting to diversify holidays effects depending on festivity (e.g. Christmas, Thanksgiving, ...) and to use these informations inside the predictive models. Anyway, the most important information that could be easily obtained are about the single markdowns. In fact, to estimate the markdowns impact, only the sum of them was used and they are not considered individually. This choice has been dictated by the poor information contained in the datasets. It would be more informative the study about types of products and relatives markdowns (applied with different ratio, e.g. 10%, 20%, 30%...). Finally, a study of elasticity parameter, considering the products of the supermarket chain object of this report, could help when the decision has to be taken.

6 Conclusions

The aim of the project was to decide if the supermarket should apply or not markdowns in a certain week of the year. From the analysis, it emerges that the most important variable to determine the final result is *Ratio* (markdowns divided by gross sales). *Weekly Sales* is another useful variable, but it depends only on *CPI* and *Unemployment* rate; it should be improved in its forecast in order to have more realistic situations. Also the variable *Temperature* is not so relevant in case of strong variance because it's a week average. The decision however is definitely based on the markdown chosen for the promotion and on ε parameter. So, at the end, a domain expert is the best person able to take that decision supported by the final model. Finally it's important to specify that, because old data concerning a particular supermarket chain has been used, the model could not give the same results if applied to different markets.

References

[1] M. Singh, "Retail Data Analytics," https://www.kaggle.com/manjeetsingh/retaildataset, 2017.

[2] R. Dalinina, "Introduction to Forecasting with ARIMA in R," https://www.datascience.com/blog/introduction-to-forecasting-with-arima-in-r-learn-data-science-tutorials, 2017.

[3] G. J. Tellis, "The price elasticity of selective demand: A meta-analysis of econometric models of sales," *Journal of Marketing Research*, vol. 25, no. 4, pp. 331–341, 1988. [Online]. Available: http://www.jstor.org/stable/3172944

[4] G. Zanni, "La domanda di prodotti agricoli ed agro-alimentari," http://www.ecostat.unical.it/anania/EMAA0809/EMAA%20La%20domanda%201%200809.pdf, 2012.

Appendix

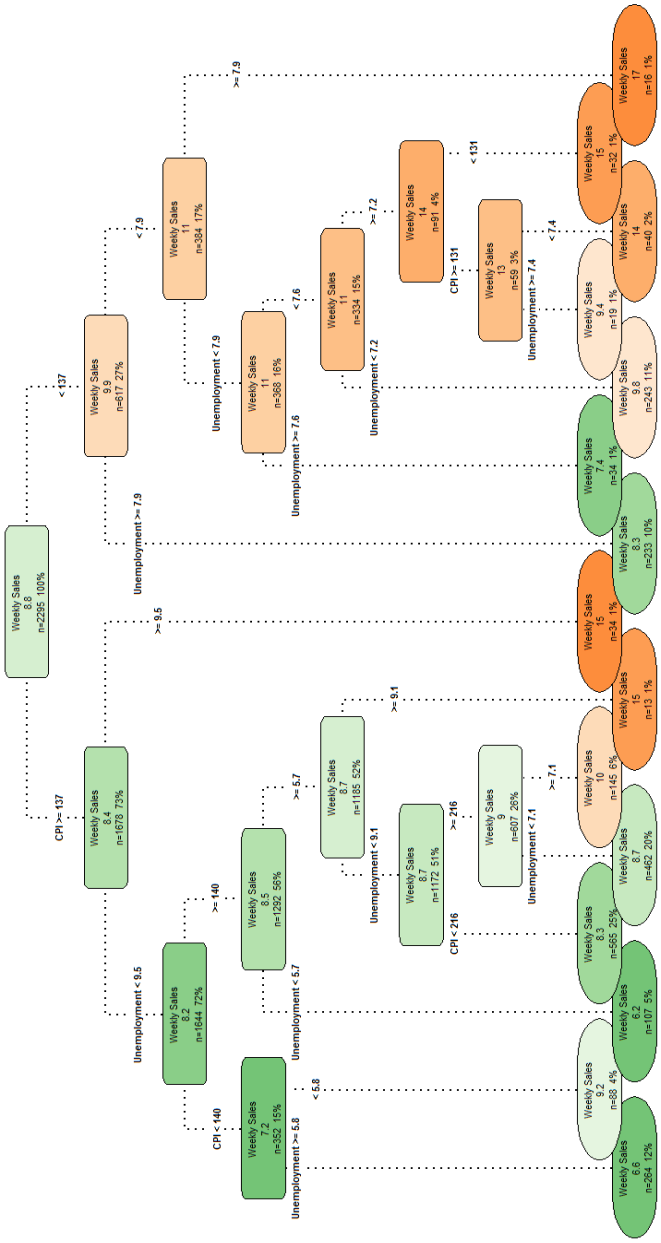

Figure 4: Tree for Weekly Sales

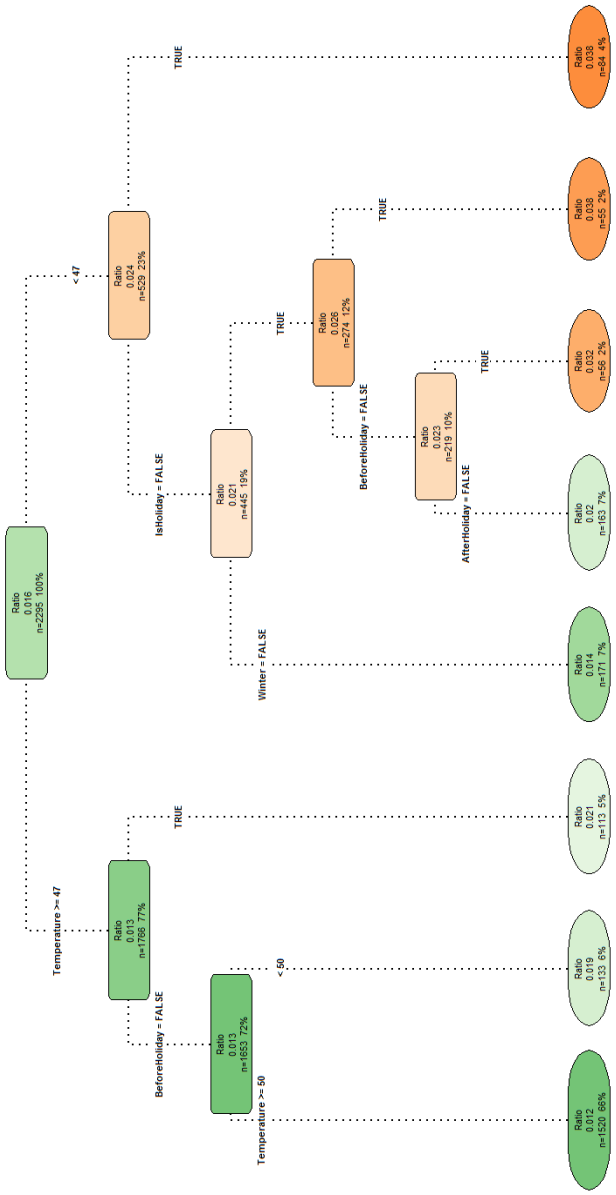

Figure 5: Tree for Ratio

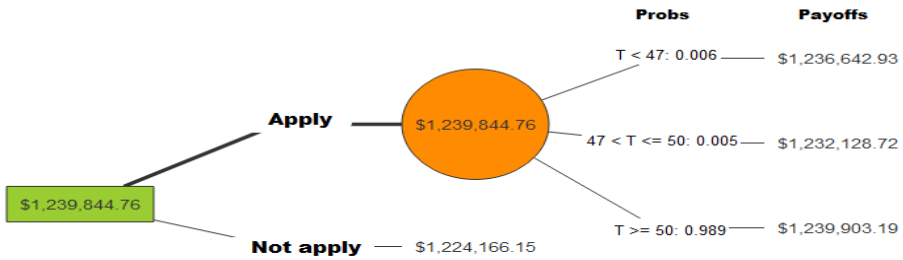

Figure 6: Decision Tree for the first week

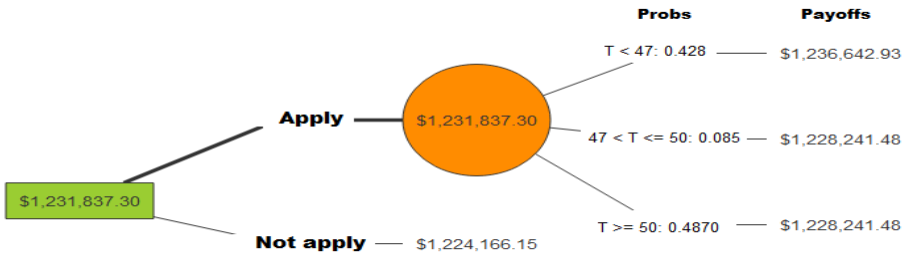

Figure 7: Decision Tree for the second week

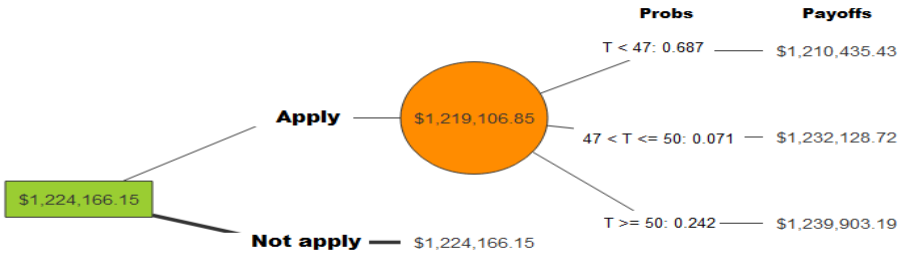

Figure 8: Decision Tree for the third week

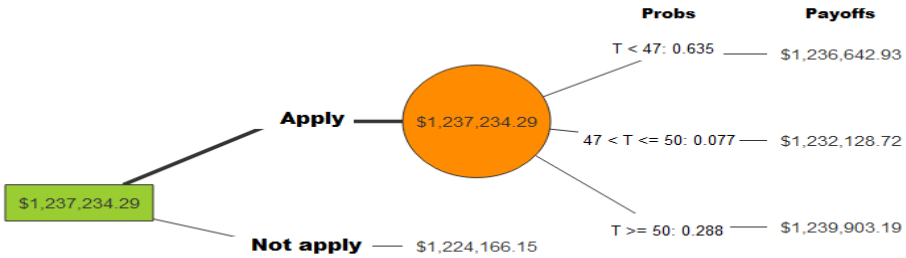

Figure 9: Decision Tree for the fourth week

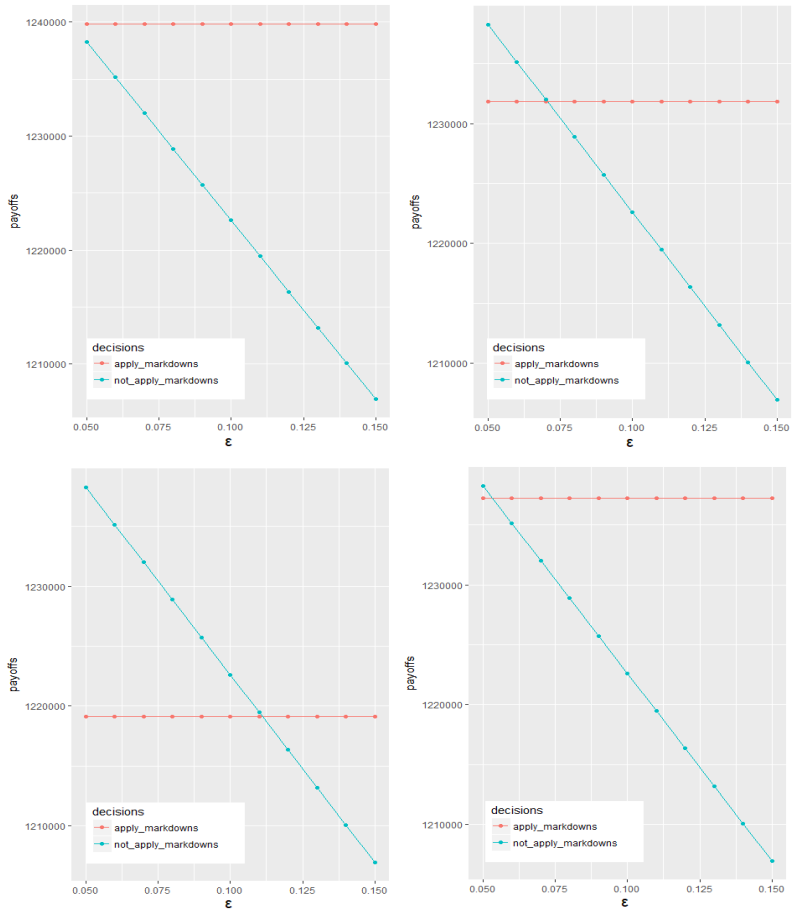

Figure 10: Sensitivity Analisys for the month