**RossmanStoreSales**

DATA SCIENCE PROJECT

Kaggle Competition: Rossman Store Sales Website: <https://www.kaggle.com/c/rossmann-store-sales> Date: September to December 2015 Autor: MSc. Oswaldo F. Domejean Email: [ofigue@gmail.com](mailto:ofigue@gmail.com) Lugar: La Paz – BOLIVIA Code: . . .

Business understanding

From the competition site description: The purpose of this competition is to forecast sales using store, promotion, and competitor data. Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied. In their first Kaggle competition, Rossmann is challenging you to predict 6 weeks of daily sales for 1,115 stores located across Germany. Reliable sales forecasts enable store managers to create effective staff schedules that increase productivity and motivation. By helping Rossmann create a robust prediction model, you will help store managers stay focused on what’s most important to them: their customers and their teams! Data

In the competition site where data is described and downloaded: <https://www.kaggle.com/c/rossmann-store-sales/data>, there is a description of every one of the features from the training and testing dataset, important to mention that the features is encoded.

Data exploration

**PromoInterval has a range of months that had been changed from ex. 'Feb,May,Aug,Nov' to 'M\_2\_5\_8\_11' and then it has been used dummy variables to create the corresponding attributes.**

The purpose of this competition was to predict the feature sales, for which the dataset has a lot of data related to the stores where those sales were made, the holydays where the store were closed, in this case depending on the region in Germany, where the stores are located spread in the country.

The dataset includes promotions and also some information related to the distance from the competition stores, and also two types of store classification the features “StoreType” and “Assortment”. And all this data registered with the date.

The date feature was divides in day, month, and year, in the case of day it was numerical, in this case it had been converted to categorical with the specific day of week.

The feature “PromoInterval” has a range of months that had been changed from ex. 'Feb,May,Aug,Nov' to 'M\_2\_5\_8\_11' and then it has been used dummy variables to create the corresponding attributes. In the case of “StoreType” and “Assortment”, it had been merged in one attribute; it showed a kind of pattern that proved to be predictive.

It had been detected that when the feature “Open” is 0 (closed Store), Sales equals 0 17% in train is with Sales equals 0. About assortment the levels “a” ad “c” have the highest sales, and it had been considered to divide the data set by using “a” and “c”, and other data set for b. It has also been identified that 54 stores have zero sales when they are open

The feature distance when it had NAs had been changed with zero, and some other features like CompetitionOpensince, Promo2SinceYear, etc., when they had NAs, it had been changed with the most common value in the feature.

Feature engineering

New features had been created like SalesCustStore where the division of sales and customer had been calculated by Store, DayOfWeek and Promo.

It also had been created SalesMean calculated by store and DayOfWeek and other features that helped in the time of the predictions.

Something that made a relevant difference were those mentioned attributes StoreType and Assortment, both individually and merged, which proved to be features that let us separate the dataset.

Models and Evaluation

The metric used to measure the accuracy was RMPSE (Root Mean Percentage Square Error) which had been identified to calculate with Cross Validation.

The techniques that had been used were XGBoost, Random Forest, Linear Regression and Glmnet. With the predictions generated it had been calculated simple mean with the result of every model created that was the ensemble process it had been used.

Conclusion

In spite of the variety of information in this dataset few features made a big impact in the prediction, two of the main ones were StoreType and Assortment. Something that draw my attention was the variety of data elements that had relevance, for example the days after the holydays the sale incremented considerably. Other one was the city in which the store resides; I mean its specific characteristic. A lot of data had impact in the predictions.