**DataMining Project**

*Team*

Flavio Mazzolini, 10493155

Pietro Menchetti, 10491926

Michael Montani, 10451817

**Data Cleaning**

*isOpen treatment*

In the training set, since for each row, isOpen = 0 implies numberOfSales = 0, we decided to eliminate each row with isOpen = 0. In this way our model learned from effectively opened stores, without learning useless and wrong relationships.

This elimination involves also the test set. We will finally sum the sales of the month for each store, predicted from each row. Eliminating rows with isOpen=0 it’s equivalent to assign 0 to the value of the sales of that specific row, which is a logic assumption.

*WindDir treatment*

For the direction of the wind, since there’s no difference of magnitude between directions, we decided to assign one over eight different winds to each angle, by dividing the round angle in 8 parts. Where the value of the angle was -1, we assigned a “wind -1” class, supposing it was an unknown or not interesting value.

*Skewness/Kurtosis analysis*

By observing the distribution of the values of some columns, we observed a relatively high Kurtosis value of its distribution. We decided though to apply log(x+1) transformation to these values, in order to obtain a distribution more similar to the Gaussian one.

We applied this transformation to precipitation and NumberOfSales, which is our target variable.

**New Features**

*Days of week*

From the date column, we decided to extract the feature “Day Of The Week”, supposing, for instance, different behaviors of sales, between the weekend and the rest of the days.

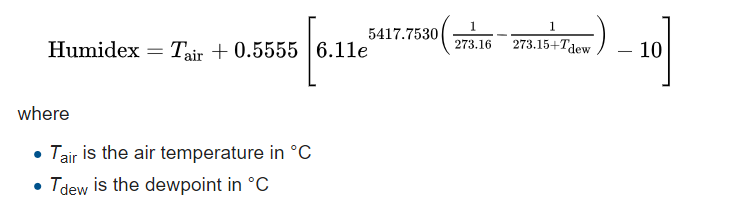
*Months*

From the date column, we decided to extract the feature “Month”, supposing to observe different sales in different months. For instance, month with different number and kind of holydays could behave differently.

*ZeroClouds feature*

CloudCover is an ordinal attribute, with integer values. By analyzing its correlation between mean sales, we observed nearly the same value of sales for each value of CloudCover, except for CloudCover = 0, which corresponds to lower mean sales. We decided though, to avoid using this value and substituting this value with an additional binary feature called “ZeroClouds”, which is equal to 1 when CloudCover = 0, 0 otherwise.

*Weather features*

Since we had a large number of weather features, we searched for other interesting features and indexes to get from the previous ones. We found H:

where

* T\_air is the temperature of the air
* T\_dew is the temperature of the dew point

It’s an index used to describe how hot the weather feels to the average person.

And then WCI:



where

* Ta is the temperature of the air
* V is the wind speed in km per hour

It’s an index used to measure the lowering of body temperature due to the passing-flow of lower-temperature air.

**OHE**

Since we had some categorical variables, we decided to encode them with One Hot Encoding. The features so encoded were:

* StoreType
* AssortmentType
* Region
* Events
* Month
* Days of week

**Missing values**

We analyzed missing values and discovered the following things:

* Max\_Gust\_SpeedKm\_h: too many missing values (78.5%), so we ended up dropping this column.
* CloudCover: since this feature becames ZeroCloud, we imputed missing values with 0, the most common situation.
* (Min, Mean, Max) Visibility: very few data were missing. We decided to impute them with the mean.

**Outliers**

By looking at the different features available we decided to look for outliers in the meteorological ones, since in our opinion are the one that, among all the others, could be affected by measurement errors. So, we plotted them with boxplots and according to their interquartile range and their quantile we decided to apply a Winsorizing transformation limiting their values to IQR + 3Q or to 1Q – IQR.

**Feature Analysis**

*Correlation Analysis*

We studied correlation between NumberOfSales and NumberOfCustomers with respect to all the other features. We used Pearson correlation coefficient to study the correlation with continuous variables, and point\_biserial coefficient for binary and one hot encoded variables.

In addition, we analized correlation of NumberOfSales and NumberOfCustomers, with squared and multiplied features, but we didn’t observe nothing interesting to add to our model.

*PCA*

We performed PCA on the continuous variables we had left, and we kept the first 10 pca. Which explained nearly the totality of the variance of the data. Waiting for other methods to do feature selection.

*Dealing with NumberOfCustomers*

Since in our train set we have “NumberOfCustomers”, but our test set doesn’t provide it, and this feature is the most correlated to our target, we ended up predict the sales in two ways:

* Predict “NumberOfSales” directly from all our features, removing “NumberOfCustomers”
* First predict “NumberOfCustomers” with a model, then predict “NumberOfSales” with a second model, taking as input the predicted “NumberOfCustomers” of the first model.

*Feature Selection*

We adopted two main methods to perform features selection:

* Lasso
* ExtraTreesRegressor

We applied these techniques for 3 kinds of models:

* The one that predict “NumberOfSales”, directly, without using “NumberOfCustomers”.
* The one that predict “NumberOfCustomers”.
* The one that predict “NumberOfSales”, given “NumberOfCustomers”.

We also applied lasso to the dataset transformed with PCA.

**Models**

*Regression*

We firstly applied simple linear regression, but we didn’t get great results, so we ended up using more powerful methods from ensemble methods.

*Crossvalidation*

In order to validate complex and heavy models obtained with ensembles techniques, we decided to perform a 3-fold crossvalidation implemented in this way:

* Shuffle dataset
* Cycle for 3 times taking as test set each successive third of the dataset
* Average r2 and RMSE of the 3 models obtained

*Gradient Boosting*

Firstly, we studied the behavior of a decision tree, and we ended up assigning 6 to the max depth of the tree. Then we performed extreme gradient boosting on a tree of that kind.

We tried to predict sales direcly without customers, and with the double model. With the features selected with lasso, ExtraTreesRegressor and lasso with pca.

The best results of r2 crossvalidated with 3 folds has been obtained by using lasso features selection, predicting sales directly without customers.

*Random Forest*

For the random forest model selection, it has been decided to analyze different performances based on different configuration of the forests and feature selection techniques. Firstly, a set of 3 random forests with 50 trees has been trained with different sets of features (all, lasso selection, random forest selection) and then the same procedure has been applied to 150 trees. The best model, according to R2 measured over the same test set, has been a model of 150 trees, with no feature selection and min\_leaf\_sample = 5.

*Final Selection*

According to our measure of 3-fold crossvalidation obtained from our ensembles models, we decided to select the model with the largest R2, which was the decision tree with 130 applications of gradient boosting, predicting sales directly without customers, with no feature selection applied, obtaining an R2 value of 0.855. We further noticed that averaging the prediction results of this model, with results of a random forest, with 150 trees, min\_leaf\_sample = 5 and no feature selection, we got a still greater r2 score (0.868). So, we finally decided to use as a final model the average between the random forest and the boosted decision tree.