Rossmann Stores Sales Prediction

Raj Nikhil Choul, Subramaniam Ganesh Kumar, Subramanian Vellaiyan, Venu Guntupalli

1870 STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Business Intelligence & Analytics

Introduction:

Å Rossmann operates over 3,000 drug stores in 7 European countries.

Å Managers art wants to predict the sales for 6 weeks in the 1,155 stores located across Germany.

Å The main aim is to build an automated robust model that predicts sales for the managers.

Data Set:

Å The dataset was downloaded from kaggle website.

Å The dataset consist of test (historical data without sales), train (historical data with sales) and store (consists of master information for the stores).

Å Train dataset has 9 columns and the store dataset has 9 columns

Methodology:

Å The train and the store data sets were merged using left join to form a master data set.

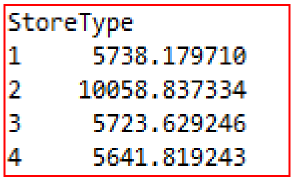
Å The join was based on the store id which uniquely identifies each store.

Å Each store’s individual data such as Store type, Assortment , Competition, Promotion etc. were mapped to each transaction in the training and test data sets.

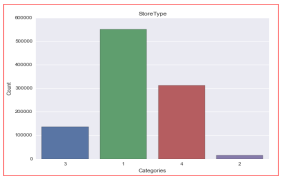
Exploratory Data Analysis :

Å Store type 2 has the maximum average sales among the store types.

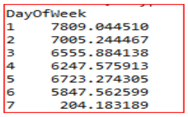
Å It is also interesting to note that the store type 2 has the least number of stores



```
StoreType
1      5738.179710
2     10058.837334
3      5723.629246
4      5641.819243
```



Å Day 1 has the maximum average sales among the Day types and 7 has the least.

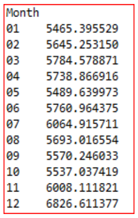


```
DayOfWeek
1      7809.044510
2      7005.244467
3      6555.884138
4      6247.575913
5      6723.274305
6      5847.562599
7       204.183189
```

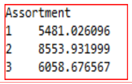
Å Months overall have same average except for spikes in July, November and December

Å Assortment type2 has the maximum average sales but implemented in fewer stores than 1 and 3

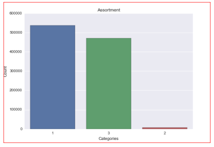
Å When there is a promotion the average salary goes up as expected

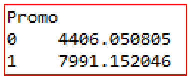


```
Month
01     5465.395529
02     5645.253150
03     5784.578871
04     5738.866916
05     5489.639973
06     5760.964375
07     6064.915711
08     5693.016554
09     5570.246033
10     5537.037419
11     6008.111821
12     6826.611377
```

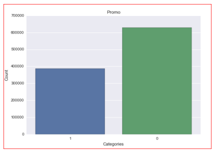


```
Assortment
1    5481.026096
2    8553.931999
3    6058.676567
```





```
Promo
0     4406.050805
1     7991.152046
```



Data Transformation:

Å The columns StateHoliday , StoreType, Assortment and PromoInterval, originally nominal variables were transformed to a dichotomous categorical variable.

Å After the transformation the categorical variables were feature engineered into new variables for the model.

Å Original StateHoliday Variable was removed and four new binary variables (SateHoliday0, StateHolidayA, StateHolidayB, StateHolidayC) were added to the model.

Å The column Date was transformed into months and weeks. Each month and week were created as a separate binary columns.

Å The missing values in the columns such as PromoInterval, which means that there are no promotions in the store currently were filled with 0s.

Model:

Å The target variable ,sales, is a continuous variable.

Å We used OLS regression , Random Forest and CART to build the model to predict the sales from August-1-2015 to September-17-2015.

Linear models:

Å Initially , linear models were built for the prediction.

Å OLS, Ridgecv and Bayesian models were implemented.

Å It is worthy to note that Bayesian has the least error with 0.40

CART:

Å We used all the featured engineered variables.

Å Store, Year Ids that were irrelevant to the model were removed form the model.

Å The CART had an error value 0f 0.19. This can be attributed to the featured engineering since it created many relevant decision rules based on the variables.

Ensemble:

Å We used two methods : Gradient Boosting and Random Forest.

Å Comparatively Random Forest performed better among the two.

Å Random Forest Had an accuracy of 0.1757 with 10 Estimators and an accuracy of 0.1751 with 25 estimators

Conclusion:

Å We noted that the variable Open determines that sale of the store.

Å Store type B together with Day 7 of Week is important for determining the sales of the store.

Å Promotion and the competition distance influences the sales of the stores.