Business Data Analytics

Homework 2

The result of your homework should be .pdf (or html) / .r / .rmd file that should contain questions and answers. In case you upload html output of rmarkdown then upload the rmarkdown as well (zip it in a folder along with rmarkdown before uploading). Use course webpage to submit the report and don’t forget to check it before submitting to ensure that vital information is readable.

The homework should contain:

- The report, which has answers to the questions along with the plots and code snippets.
- In order to submit these two files, they need to be zipped first.

You are always welcome to ask us any questions in Piazza! With the gained knowledge during the lecture and lab session you should be able to solve the following tasks:

1. (2 points) Using orders_rfm.csv data, perform the RFM analysis similar to what we did in practice session. But let the reporting date be “2017-04-01” (Year-Month-Date format). Divide clients into bins using quantiles. Interpret the resulting clusters. Compare with the results that we got during the practice. Was the result better or worse? Why?

2. (4 points in total) Imagine that you are working in a company that offers different types of furniture. Use the customer survey data “customer_survey.csv” that contains information on customers. Columns:

- age: age of the customer
- gender: male or female
- income: total income of the customer
- kids: number of kids
- ownHome: whether customer owns house or not
- subscribe: whether customer subscribed to mailing service or not

Execute the following tasks:

2.1 (1 point) Perform data transformation on the existing dataset:

a. Convert all factor features to numerical. Each different value should be converted to separate number. Example: Yes => 1; No => 0.

b. As a next step, rescale all features.
2.2 (2 points) Use Elbow method to determinate optimal amount of clusters for k-means clustering. For each amount of clusters you take use at least 10 different initial centers. Report the most optimal number of clusters.

2.3 (1 points) Pick up 30 records randomly from the dataset and perform hierarchical clustering and build a dendrogram based on this 30 records. Pick the first four clusters (based on similarity) and cross check their feature values in the dataset. Does it make sense that the first two clusters are more similar (in terms of features) than the next two?

3. “HousePrices.csv” contains information about prices for different types of houses. The list of the columns:

- Area – id of area where house is placed;
- Garage – amount of garages in the house;
- FirePlace – amount of fireplaces;
- Baths – amount of baths in the house;
- White Marble – whether house has white marble or not;
- Black Marble – whether house has black marble or not;
- Indian Marble – whether house has black marble or not;
- Floors – amount of floors in the house;
- City – id of the city in which house is placed;
- Solar – whether house is solar or not;
- Electric – whether there is electricity in the house;
- Fiber – whether fiber is used;
- Glass Doors – whether doors created from the glass;
- Swimming Pool – whether there is swimming pool or not;
- Garden – whether there is garder;
- Prices – price of the house.

Use this data for training the model which can predict prices.

3.1 Prepare dataset for prediction. Split data into:

   i. Case A: 60% of train data & 40% test
   ii. Case B: 75% of train data & 25% test

3.2 (1 point) Calculate correlation between price and each feature. Which are top 8 features that has highest correlation with price?
3.3 (2 points) Build the models using the training data for each of the cases specified above (Case A and Case B) and then predict for the test data. Select features (according to your choice) so that RMSE of the model is not more than 2000.

3.4 (2 points) Build the table. For each model that you have trained report:

- Time that was taken to train the model;
- RMSE;
- MAE;

<table>
<thead>
<tr>
<th>Case</th>
<th>Feature names</th>
<th>Time</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.5 (1 point) Analyze the results and report which model is the best for price prediction? (For this task you don’t have to code, just use the table above to give your reasoning for your answer).