CMPT 741- Course project
Fall 2019

Predict Future Sales

This challenge serves as a course project for the CMPT 741 course, which has been designed based on the final project of "How to win a data science competition" Coursera course. In this competitional project. (see https://www.kaggle.com/c/competitive-data-science-predict-future-sales for more information)

you will work with a challenging temporal dataset consisting of daily sales data, kindly provided by one of the largest Russian software firms - 1C Company. The main goal of this project is predicting total sales for every product and store in the next month. However, there are some intermediate stages which enhance your analytical skills and prepare you for the final goal.

- **Each group has to sign up for the Kaggle competition "Predict Future Sales" and provide us with their team information**

Phase1. Exploratory data analysis and data cleaning

Like all the data mining and machine learning pipelines, we expect you to perform the initial analysis and exploration on the dataset to summarize its main characteristics. This step is a great practice to see what the data can tell you beyond the formal modeling or hypothesis testing task, like discovering the potential patterns, spotting outliers and so on. To this aim, you can apply any arbitrary, but meaningful visualization and statistical tests. It is worth noting, that you may incrementally improve your exploration section, as you work on the next steps.

Obviously, in this phase, you need to perform data cleansing, which is the practice of detecting and correcting corrupt or inaccurate records from the dataset, by identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Feel free to apply your arbitrary data preprocessing as long as they are insightful and maximize the accuracy without necessarily deleting information.

Phase2. Feature engineering

The second critical step of your pipeline is supposed to be feature engineering, which is the process of extracting features from a raw dataset. You need to refine the available data into features, predictor variables, and add date-based predictors to feed the final methods with purified oil. Without relevant features, you can't train an accurate model, no matter how complex the algorithm. Obviously, this step is highly dependent on the model that you are aiming to apply on the dataset.

Due date	October 8, 2019
Presentation	Submission + presentation in office hour

Phase3. Clustering

We expect you to cluster shops and item categories based on the available transactions. Feel free to use any type of clustering method, as long as you can justify why that method(s) suits your dataset and goal. How can the result of clustering help you to perform the final prediction task?

Due date	October 22, 2019
Presentation	Submission + presentation in office hour

Phase4. Design and Implementation of Predictive Models

Based on the collected information and engineered features, the final justified methods should be designed and implemented. We expect each group to apply two types of methods on the dataset:

- A standard model. Designing any arbitrary but suitable variations of standard method on the dataset, to perform the forecasting task. Please be noted that including the date-based engineered features may hugely boost your model performance.
- A time series mode. Designing and implementing an appropriate time aware regressor model.

Due date	November 5, 2019
Presentation	Submission

Phase5. Train and Test of Models

Last but not least, perform the training step and apply the required improvements. Using the online testing process, evaluate your model performance and improve it. **To evaluate the performance of your models, you need to apply the online testing process of Kaggle.**

Phase6. Project Poster and Report

At the end of the semester, each group is supposed to prepare a comprehensive report and submit it plus their source code or the git link and access. Moreover, we will have a poster session presentations to provide the group the opportunity of presenting their findings and results.
The project poster and the report shall consist of the following sections: 1. Introduction, 2. Feature Engineering, 3. Clustering, 4. Design and Implementation of Predictive Models 5. Experimental Results, 6. Conclusion. The poster and the report will be evaluated according to the following three criteria: originality, technical quality (includes the prediction accuracy as obtained from the Kaggle test), and clarity of presentation.

Due date (Poster)	November 26, 2019
Due date (Report)	November 30, 2019
Presentation	Submission + Poster presentation