Data Validation and Anomaly Detection for Shiny Dashboard in R

Capstone Project Summary

Master of Science in Business Analytics

Aron Palkovics

2020

Capstone Project Summary

Business Analytics

The project holder of the capstone project is an Independent Hotel Operator company in South-East Asia.

As Data Scientist say 80-90% of their work is cleaning/organizing and validating data, looking for extreme values, anomalies, duplicates, missing values and much more. This is essential to ensure reliable results.

The project sponsor's data consists of numerous tables that needs to be validated in order to analyze historical data to facilitate business decision making, to create interactive functions based on the needs of the different departments and to automate reporting accurately. The company is currently in the process of developing a Validation Tab on their Shiny Dashboard. The data can be very noisy, often containing missing values, duplicates, errors (due to manual input). The company store 10s of millions of records on over 150 hotels, receiving financial and other reports on daily bases. It's essential for them to create an automated easily readable system to ensure data quality is acceptable.

To fulfill the requirements I deployed an easily reproducible pipeline in R which is composed of many data quality validator and anomaly detector tools as the following:

The data was presented in parquet file format, the first step was a basic plausibility and validity check performed on all data sources before relevant features and subgroups of the table was collected. I used several methods which gives automated, summary and filtering options for finding observations with quality errors, missing values (more filtering option for rows with only one or more missing features), duplicates (the data is built on of daily observation basis, sometimes more times updated during the day) and created some features, which captures seasonality in the time series data.

The main part of the project contains the analyzation which, is a mixture of statistical models, and previously deployed R packages, for detecting any extreme value in the given time series on the given decision threshold. The pipeline working for hotels individually filtered by hotel ID (the goal was the total interactivity on the dashboard, all the relevant input options are represented).

- Grubbs' test
- K-nearest neighbors distance method (both pairwise and on all numeric standardized features)
- LOF Local Outlier Factor for local and global anomalies with the metric of Gower distance to implement distance across observations with qualitative and quantitative features (seasonality factors)
- Anomaly Detection using Seasonal Hybrid Extreme Studentized Deviate test and algorithm, decomposition is used to remove trend and seasonal components
- Benford's low for mathematically non outlying observations
- And some more which are also deployed or just recommended in the technical documentation

As a final outcome, an interactive multi plot object and an interactive data table was created where the different anomaly detection methods can be combined. The plot shows the found extreme values on the regular data for the given hotel in the given timeframe and capable of zooming in for specific timelines and data sectors (synchronized for each feature plot). From the table, interactive filtering, searching, data representation and buttons for copying, printing, downloading the required data in csv, excel and pdf is given. Both these metrics are built as a pipeline of functions which makes the deployment clear for a Shiny Dashboard.

To sum up, the previously discussed project pipeline is capable for many different kind of data validation methods, as a non-exhaustive list of the many more possible opportunities, while it is consummate for saving time and energy. The found data points can be easily analyzed manually or by an automated system handled by the project holder company's data science team.