CSP 571 - Data Preparation and Analysis
Course Project
Fall 2018

0 Project Group Members and Area:

Area: Business sector (customer behavior relationship with spent amount).

- Muhammad Zubair Shareef (A20303179) - Leader
- Abdulwahab Alothaim (A20412364)

1 Project proposal

1.1 Introduction

The expansion of the internet and the increasing percentage of online purchases transactions drove many companies in the retail business sector to work to build their e-commerce websites' stores. Any action on the website could be tracked, recorded, and analyzed with the aim of understanding customers' behaviors and the connection with the customers' purchases.

1.2 Research Goal

1.2.1 Description

Typically, a small percentage of customers produce most of the revenue for a business, and marketing teams need to identify these customers in order to create effective marketing strategies. Our objective is to create a predictive model for expected revenue per customer in the Google Merchandise Store, in order to identify top customers. This information can then be used to optimize marketing strategies to increase sales.

1.2.2 Specific Questions:

- What impact do the features of the data, such as device used, geographic location, and date and time have on the total revenue for a transaction?
- Which features have the strongest effect on the prediction?
- Which features have the most effect on customers' leakage?

1.2.3 Our prediction

Our goal is to predict the natural log of the sum of all transactions per user. For every user in the test set, the target is:

$$y_{user} = \sum_{i=1}^{n} transaction_{user_i}$$

$$target_{user} = \ln(y_{user} + 1)$$

The final prediction model uses log(revenue) as the only response, and not simply revenue. Using the log makes it so the problem is more about predicting the order of magnitude of revenue from a customer, instead of a precise value. This makes sense as a business goal as businesses are interested in predicting the revenue so that they then can identify and target high value customers. It would be less useful, and probably more difficult, to train a model to exactly predict the value.

1.3 Methodology

Our methodology consists of three main parts:

1.3.1 Data Preparation:

- Aim to dissolve the complexity of the dataset and prepare it for the next step.
- Our data contains a mixture of unavailable data and nested fields. Hence, we need to dissociate variables and reunion them again.
- The output of this step are two data frames; one for revenue training data, and the other for customers' leakage.

1.3.2 Data Discovery:

- Aim to discover our dataset to help us efficiently and optimally process next steps.
- We need to go deep inside our dataset and represent each feature and extract the five number summary in order to make our decision about data cleaning, wrangling, and modeling.
- The output of this step are some facts about each feature in our dataset.

1.3.3 Data Wrangling and Modeling

- Aim to reach the goal of this project, and answer the main questions we have.
- We may clean, rearrange, and split data in order to find the best model that fit the data, and estimate the revenue for each customer.
- The output of this step is a model that has the least root mean square error (RMSE)

  $$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

  where y hat is the natural log of the predicted revenue for a customer and y is the natural log of the actual summed revenue value plus one.
- Test the model on the test data set, and adjust it as necessary to improve the accuracy

2 Project Outline

2.1 Literature Review

- Resources identified in the reference section need to be reviewed in order to gain background knowledge of the problem.

2.2 Data Source and References

2.2.1 Overview

The dataset is from a Kaggle competition for the Google Merchandise Store, and it contains both training and test customer transaction information. The training dataset has 903,654 rows and the test dataset has 804,685 rows, both data sets have 12 columns.

- Four of the columns (device, geoNetwork, totals, and traffic source) are JSON columns, so they contain information for multiple columns within them. These columns need to be flattened.

2.2.2 Feature Description

- The details of the dataset, from Kaggle, are explained below:

Field Name	**Type**	**Description**
fullVisitorId	String	A unique identifier for each user of the Google Merchandise Store.
channelGrouping	String	The channel via which the user came to the Store
date	Date	The date on which the user visited the store
Device	JSON	The specifications for the device used to access the Store
geoNetwork	JSON	This section contains information about the geography of the user
sessionId	String	A unique identifier for this visit to the store

socialEngagementType	String	Engagement type, either "Socially Engaged" or "Not Socially Engaged".
totals	JSON	This section contains aggregate values across the session, such as visits, hits, pageviews, totalRevenue
trafficSource	JSON	This section contains information about the Traffic Source from which the session originated.
visitId	String	An identifier for this session. This is part of the value usually stored as the _utmb cookie. This is only unique to the user. For a completely unique ID, combination of fullVisitorId and visitId needs to be used.
visitNumber	int	The session number for this user. If this is the first session, then this is set to 1.
visitStartTime	POSIX Time	The timestamp (expressed as POSIX time).

2.3 Data processing

- Some of the columns are JSON objects, which contain information for multiple features. These JSON columns need to be flattened.
- Some of the flattened JSON columns do not contain any important information, as the value is a constant for the entire dataset.
- Some of the categorical features have extremely high cardinality, such as the geoNetwork column. This can be addressed by using one hot encoding, label encoding.
- Duplicates need to be searched for and removed.

2.4 Model selection

- Our project consists of two parts; revenue prediction for an observation and leakage detection.
- For the first one we are going to use a regression model with Regularization because we have a large number of variables, and most of them are not independent of each other. Also, we want to ensure that no feature drags the model to its side and make our model more bias.

- Regarding the second part, we are going to use a two class logistic regression where the observation will be classified as a leak or steady customer.
- However, RMSE have a huge impact of this process. Therefore, we may change our model in the future.

2.5 Software packages and Dataset

- We will use RStudio for all steps of this project, including data preprocessing, data cleaning, and data analysis. The goal is to create a model that predicts the natural log of total revenue per customer in order to identify the segment of customers that will generate the most revenue.
- Libraries: ggplot2

- **Resource**: Kaggle
- **Dataset**: Google Merchandise Store[1] Customers' information and transactions.
- **Description**: Each record in the dataset has a set of customer information and the revenue that was collected from this customer based on one or multiple orders. The details of the dataset were explained in 2.2.2

3 Reference Resources:

- Yip, Patrick (2003). *U.S. Patent No. US20040236649A1*. Washington, DC: U.S. Patent and Trademark Office.
    - Customer revenue prediction method and system. A patent by Google which consists of a method for predicting revenue associated with an account, for a specific time period, based on historical revenue data.
- **Academic Papers:**
    - Eugene Wong, Yan Wei, (2018) "Customer online shopping experience data analytics: Integrated customer segmentation and customised services prediction model", International Journal of Retail & Distribution Management, Vol. 46 Issue: 4, pp.406-420, https://doi.org/10.1108/IJRDM-06-2017-0130
        - Authors segmented high-value customers, analysed their online purchasing behaviour and predicted their next purchases from an online air travel corporation.
- **Reference Books**:
  We will use these to understand the underlying statistical techniques in order to construct our model
    - *Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013. Print.*
    - *Hastie, T., Tibshirani, R.,, Friedman, J. (2001). The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc..*
    - *Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. New York: Springer.*