

SMU Customer Purchase

Behavior Capstone

Wednesday, 08 May 2019



**Table of Contents**

Table of Contents

[Goal: 2](#_Toc8313251)

[Customer: 2](#_Toc8313252)

[Project Team: 2](#_Toc8313253)

[Advisors: 2](#_Toc8313254)

[Methodology: 2](#_Toc8313255)

[Tool Recommendations: 3](#_Toc8313256)

[Set-up: 3](#_Toc8313257)

[Tips: 3](#_Toc8313258)

[Data Dictionary: 3](#_Toc8313259)

[Deliverables: 4](#_Toc8313260)

# Goal:

As an International Manufacturing Company, we have a very large and diverse customer base which spans every continent as well as the following industries: Oil and Gas, Chemical, Power, Water and General Industry (e.g. Agriculture, Mining and many others). It’s essential that we do everything necessary within financial and ethical means to meet their needs. One way we are looking to do so is to leverage our Enterprises global bookings data to understand purchase behaviors within our aftermarket operations – specifically for parts, which are used to repair and maintain equipment operating in the field. Our problem is three-fold in that we would like for you to analyze our customer bookings data repository for the following: 1) Associations within and across various part product families and customers, 2) Predict future purchases of various parts for customers, and 3) Understand the propensity of the customer to churn, and when. You will be given significant freedom for your analysis and predictions.

A Non-Disclosure Agreement (NDA) is required to participate in the project

# Customer:

Flowserve Transformation Analytics Center of Excellence Team

# Project Team:

SMU Capstone Team

# Advisors:

Brent Allen, Flowserve Data Scientist

# Methodology:

I recommend the Cross-Industry Standard Process for Data Mining (CRISP-DM) model is used:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Visualization
5. Deployment
6. User Acceptance
7. Documentation

# Tool Recommendations:

Below is a brief list of recommended tools to use for your analysis:

Anaconda python distribution -> <https://www.anaconda.com/distribution/>

(Spyder and Jupyter are included with Anaconda)

Git -> <https://git-scm.com/downloads>

GitHub -> <https://github.com/>

VSCode -> <https://code.visualstudio.com/download>

Pycharm -> <https://www.jetbrains.com/pycharm/download/>

Sublime -> <https://www.sublimetext.com/3>

# Set-up:

In using Anaconda, set-up your environment with an environment file. An example is provided for a basic python 3 environment. The Anaconda environment can be created with the following command:

**conda env create -f env\_custbb.yml**

You must activate the environment to use it. The following code will:

windows -> **activate custbb**

linux -> **conda activate custbb**

Any Integrated Development Environment (IDE) will work for this project that integrates with python. My preference is VSCode but any IDE will work.

# Tips:

1. When loading the data into a pandas DataFrame ensure you use the encoding ‘latin-1’
2. A useful package is scikit-survival which is utilized the cox proportional hazards algorithm and can be found at <https://github.com/sebp/scikit-survival>.
3. Critical to this project will be how you define “Customer Churn”

# Data Dictionary:

There are five files each of which is ~500MB. These files contain part orders from January 01, 2014 to May 8, 2019. Each file contains the following columns:

'Date' -> the date of the order in the format ‘YYYY-MM-DD’

'OrderNumber' -> order number in an alpha-number format

'OrderLineNumber' -> the line number of the order

'Quantity' -> the quantity of the part ordered

'SellPrice' -> price the part is sold in Dollars

'TotalOrderCost' -> total cost of the order as identified by “OrderNumber” in Dollars

'CCN' -> Coporate Control Number; a part identification number at the corporate level

'PartnerNumber' -> the part number at the plant level

'Material' -> material identifier for the part

'ItemCategoryGroup' -> parent category of “ItemCategory”, group of the part sold

'ItemCategory' -> sub category of “ItemCategoryGroup”, a more specific description of the part

'ItemDescription' -> text description of the part

'ProductCode' -> a high level product code of the part; this could be a pump, valve or seal

'ProductDescription' -> text description of the “ProductCode”

'NounCodeDescription' -> text description of the “NounCode”

'NounCode' -> a part category which is different from “ItemCategory” and “ItemCategoryGroup”

'ShipFrom' -> the Flowserve plant the part was shipped from

'CustomerName' -> the name of the customer

'CustId' -> alpha numeric customer identification

'Address' -> street address of the customer

'City' -> city of the customer

'State' -> state of the customer

'PostalCode' -> postal code or zip code of the customer

'Country' -> country of the customer

'Region' -> region code of the customer; geographic region

'SubRegion' -> child category of “Region”; more specific geographic region

'Industry' -> industry code of the customer; sub category of “IndustryGroup”

'IndustryGroup' -> high level category of customer; parent category of “Industry”

'Plant' -> Flowserve plant the part was produced

'PlantType' -> classification of the plant

'PlantPlatform' -> type of product of the plant produces; related to “Product Code”

'Market’ -> market classification of the customer

# Deliverables:

1. Code used to develop your analysis
2. Copy of capstone write-up
3. PowerPoint presentation of your analysis limited to 10 slides total