Bellevue University

DSC630 Predictive Analytics

Final Project Proposal

**Team**

Gangadhar Dhulipala

Anirban Pal

Raghu Raman Nanduri

6/11/2019

## **Introduction**

### **1.** **Project Background**

|  |  |
| --- | --- |
|  | The project is a part of the final submission for the Summer 2019 Predictive Analytics class (DSC630) as part of the MS Data Science program in Bellevue University, Nebraska. We would be predicting the outcome of a specific scenario, based on open data available in the internet, by applying predictive modelling and other machine learning concepts.  Currently, together we selected 3 datasets from our independent research. Within the next week or so we will pick 1 out the 3 datasets that are detailed below as Option1, Option 2 and Option 3. During the selection process we will get an opportunity to weigh out the pros and cons of each others dataset and come up with a single dataset that has more scope to be a final project.  **Option 1:**  Dataset Name: LA Metro Bike Share  Data Source: <https://bikeshare.metro.net/about/data/>  Metro Bike Share is a bicycle sharing system in the city of Los Angeles, California Metropolitan area. The bike sharing system was launched in 2016. It is administered by the Los Angeles County Metropolitan Transportation Authority and is operated by Bicycle Transit Systems.  The reason this topic is selected is because there are some eminent positive positive effects of bike sharing related to environmental sustainability and the personal health benefits experienced by users. Bike share schemes like Metro (LA), Pronto (Seattle) etc. are intended to “greening” of the city; ideally, more bikes can equal fewer cars, which leads to a reduction in greenhouse gases and an improvement in local air quality. It also has significant health benefits.  **Option 2:**  Dataset Name: Online Retail Dataset  Dataset Source: <https://archive.ics.uci.edu/ml/datasets/Online+Retail>  Dataset Summary: This dataset is from UCI machine learning library. The dataset contains information about online retail store. The information includes retail data corresponding to an online retail store which includes a total of 8 attributes and 541K instances. The columns in the dataset include, Invoice Number, StockCode, Description About the Item, Quantity, Invoice Date, Unit Price, Customer ID and Country.The reason for the topic selection is that it has a very huge scope to perform a wide variety of time series analysis. As a part of this time series analysis, we can take advantage of different visualizations that can offer insights about the data. Further, the attributes included in the dataset has a mix of attributes of different data types. Between all the things mentioned we will have an opportunity to improve our data cleaning and data wrangling methods.  **Option 3:**  Dataset Name: Inpatient Prospective Payment System  Dataset Source: <https://data.cms.gov/Medicare-Inpatient/Inpatient-Prospective-Payment-System-IPPS-Provider/fm2n-hjj6>  Dataset Summary: Inpatient Utilization and Payment Public Use File (Inpatient PUF) - data summarizes the utilization, medicare payments of inpatients for the various Medicare Severity Diagnosis Related Group (MS-DRG) and covers Fiscal Year (FY) 2016. The data was made public by the The Centers for Medicare & Medicaid Services (CMS). This Inpatient dataset contains hospital-specific charges for the more than 3,000 U.S. hospitals that receive Medicare Inpatient Prospective Payment System (IPPS) payments paid under Medicare based on a rate per discharge using the Medicare Severity Diagnosis Related Group (MS-DRG). The data is aggregated at the provider identifier and Medicare Severity Diagnosis Related Group (MS-DRG). Want to expand the scope by looking at related data sets available on the “Centers for Medicare & Medicaid Services’ website - cms.gov. |

### 

### **2.** **Problem Statement**

|  |  |
| --- | --- |
|  | **Option 1**  Using the LA Bike Share data try to predict “Pricing plan maximizing total revenue from bike usage” based on tickets and pass-types, trip duration, and number of trips by region. If possible, create a mobility modelling based on the historical bike usage data.  **Option 2:**  The goal of the Online retail dataset is to identify different groups of customers by applying clustering based algorithms. After, applying the clustering algorithms and applying the labels, we can further perform the classification algorithms to classify the dataset according to the assigned labels.  **Option 3:**  The idea behind selecting this Medicaid & Medicare data set as one of the options is to analyze to find patterns and forecast the diagnosis costs for various illnesses that are part of the MS-DRG and relate it the various raising costs of insurance. |

### 

### **3.** **Project Scope**

|  |  |
| --- | --- |
|  | **Option 1**  Scope of the project is limited to data discovery, data acquisition, data quality, data preparation, data assumptions, exploratory data analysis, predictive model creation, visual representation and final project report creation.  **Option 2:**  Coming to the scope of the project, it gives us the required practice with respect to data preprocessing techniques, performing time series analysis and visualizations. In addition, given the primary question is to group the customers by applying the clustering methods, the dataset will be further classified based on the labels we get from the clustering algorithm  **Option 3:**  Scope of the project involves analyzing the available CMS - Inpatient data set, mining and merging data from other data sources on CMS to come up with a complete analysis on various Diagnosis related medicare costs for inpatients. Analysis includes visualizing the trends in the covered charges & payments for different illnesses in different states to identify costly diagnosis and costly states. Create a grouping of the diagnosis groups based on the costs incurred for the inpatients using clustering algorithms. Predict/ Forecast costs and payments for other diagnosis using Regression algorithms. |

## 

## **Preliminary requirement**

## **Technical approach**

### **1.** **Analysis**

|  |  |
| --- | --- |
|  | **Option1**  Analysis will consist of data discovery and exploration. Specifically,   * Finding the right data source * Creating project structure * Data analysis, munging, transformation * Exploratory data analysis * Predictive Analytics   **Option 2:**  For the analysis part, there are two major components:   * Data Pre-Processing * Clustering and Classification Algorithm   **Option 3:**  Approach involves the following phases   * Data cleansing * Data merging & Transformations * Exploratory data analysis & creating visualizations * Clustering and Regression Algorithms |

### **2.** **Requirement Development**

|  |  |
| --- | --- |
|  | **Option 1**  Business and Functional Requirement:  Metro Bike Share company needs an effective way to stuff it’s bike and roll out new membership or continue existing membership so that their revenue is maximized and high level of customer satisfaction is achieved through membership offerings and bike availability.  Technical Requirement:  Three basic datasets are required for the model. They are:   1. Bike Share trip data with duration, geocoding, and membership type. 2. Station details with code and location.   **Option 2:**  The goal of the retail company is to group its customer base into different clusters, so that we can segregate the customer base according to the spending power. In addition, the analysis can be used to address several other questions such as which country has the biggest and lowest customer base, what time of the week is good for selling a certain product and several other similar questions can be answered.  The advantage with this dataset is that all the data available as a single large file. Hence, no major timeless with respect to the data discovery. However, the only disadvantage with this project is that it doesn't give much opportunity to perform a lot joints and unions.  **Option 3:**  The requirement of the analysis is the find out the patterns between various diagnosis related illness and the inpatients charges and payments for those diagnosis. In order to see the complete picture there might be need to find and merge other data sources like CMS provider of service’s data and that’s the challenge in this data set. |

### 

### **3.** **Model Deployment**

|  |  |
| --- | --- |
|  | **Option 1**  The scope of this project is to build and evaluate the model. Deployment is out of scope, since we do not have a cloud solution for deployment.  **Option 2:**  The project will end by building and evaluating the models i.e. both clustering and classification models. Hence, no deployment is included within the scope of the project.  **Option 3:**  Project involves just building a forecasting model and clustering different groups of diagnosis. Deployment is out of scope as we do not have the future features related to it. |

### 

### **4.** **Testing and Evaluation**

|  |  |
| --- | --- |
|  | **Option 1**  The model should be tested and evaluated based on the test and train data randomly sampled from the project dataset.  **Option 2:**  The models built in the project will be evaluated by 3 fold cross validation. The average accuracy will be used as a metric to determine the performance of the project  **Option 3:**  The clustering model build will be evaluated using metrics using distance between the points in the clusters with respective to the cluster centers. That should help is creating the groups for various diagnosis. |

### **5.** **Specific Exclusions from Scope**

|  |  |
| --- | --- |
|  | **Option 1**  Depending on the processing performance, we may or may not include all trip data since 2016.  **Option 2:**  Since the dataset has only 8 attributes and 541K rows, we don't anticipate any exclusions. The only time we might consider exclusions in the data would be when we come across some missing data and outliers.  **Option 3:**  The data set is aggregated at the provider identifier and Medicare Severity Diagnosis Related Group (MS- DRG). The data set does not include the per patient related information and also does not include the hospital's entire population, it only represents inpatient charges and payments for Medicare beneficiaries with Part A fee-for-service coverage. |

## **Expected result**

|  |  |
| --- | --- |
|  | **Option 1**  Once finished, the project will tell us which membership type has potential to earn the most revenue for the bike share company, as well as which route can have the maximum traffic. This will help the bike share company to maximize its revenue, and optimize its bike stuffing in different stations.  **Option 2**  At the end of the project we will be able to group the total number of customers into certain groups based on the clustering algorithms. By using this information, the retail store will have a better idea about the attributes that will help them identify the customers with high spending power.  **Option 3:**  If we choose this data set, the expectation from this project is to find trends in inpatients costs and group various diagnosis and providers from different states and also predicting the future inpatient charges for those diagnosis. |

## **Management approach**

### **1.** **Project Plan**

|  |  |
| --- | --- |
|  | **Option 1**  The project will be released in an incremental MVP (minimal viable product) approach as the project is phased as assignments over few weeks. The weekly assignment also governs what will be released each week.  **Option 2:**  The project will be built in several increments within the time bounds specified by the final project of this course  **Option 3:**  The project will follow the CRISP-DM methodology as much as possible to deliver the best results. |

### **2.** **Project Risk**

|  |  |
| --- | --- |
|  | **Option 1**  Since we would be working on this project along with other major assignments and the summer term is shorter, timeline is one risk that we need to mitigate. Collaboration from different geographies pertains the risk of not having effective communication and collaboration. For the project, would be collecting open data from internet, hence data quality might be a risk. During the analysis, if we need any supplemental data, finding that data might be a risk, given there is no central data source.  **Option 2:**  It would have been much better if there are few more attributes in the dataset so that the algorithm has a bigger scope of identifying the most influential parameters that has the ability to segregate the customer base. However, there are only 8 attributes. This is the risk we are taking at this point if we were to choose this as our final dataset for our project.  **Option 3:**  To perform the expected predictive analysis, it is imperative to find other ‘Centers for Medicare & Medicaid Services’ data that could be merged with this data set to let us see the big picture. The risk involves finding such data set from the CMS website. |

## **Approval and Authority to Proceed**

We approve the project as described above, and authorize the team to proceed.

|  |  |  |
| --- | --- | --- |
| **Name** | **Title** | **Date** |
| Anirban Pal | Member | 6/11/2019 |
| Gangadhar Dhulipala | Member | 6/16/2019 |
| Raghu Raman Nanduri | Member | 6/16/2019 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |