Team V (Victory)
Pinmanee Eowpittayakul
Yawen Han
Shangyao Liu
Jincheng Xu

Final Project: Checkpoint 2

Dataset

The "Telco Customer Churn" dataset was downloaded from IBM Watson Analytics. The goal is to predict customer churn rate and gain understanding on customer behaviors so that the company can better retain them. For this project, our team will use this dataset to analyze customer behaviors, with the goal to develop a dashboard that sales and marketing team can use to understand customers and develop customer retention programs.

Dataset Link:
https://community.watsonanalytics.com/wp-content/uploads/2015/03/WA_Fn-UseC_-IT-Help-Desk.csv

Data Description

Churn happens when a customer stop doing business with a company. Telecommunications industry is one of the industries that particularly interested in the churn rate since customers usually have a relatively long contract with the companies and the customers have multiple choices. This dataset contains 7043 rows (customers) and 21 columns (features). The "Churn" column is our target. Most of our data are categorical like Churn, Gender, Dependents, Multiple Lines, Internet Service, Contract, Payment Method, etc. And the numerical data are like Tenure, Monthly Charges, Total Charges.

Variables Description

Variables	Description
Customer ID	Customer ID
Gender	Whether the customer is a male or a female
Senior Citizen	Whether the customer is a senior citizen or not
Partner	Whether the customer has a partner or not
Dependents	Whether the customer has dependents or not
Tenure	Number of months the customer has stayed with the company
Phone Service	Whether the customer has a phone service or not
Multiple Lines	Whether the customer has multiple lines or not
Internet Service	Customer's internet service provider
Online Security	Whether the customer has online security or not
Online Backup	Whether the customer has online backup or not
Device Protection	Whether the customer has device protection or not
Team Support	Whether the customer has tech support or not
Streaming TV	Whether the customer has streaming TV or not
Streaming Movies	Whether the customer has streaming movies or not
Contract	The contract term of the customer
Paperless Billing	Whether the customer has paperless billing or not
Payment Method	The customer's payment method
Monthly Charges	The amount charged to the customer monthly
Total Charges	The total amount charged to the customer
Churn	Whether the customer churned or not

Team V (Victory)
Pinmanee Eowpittayakul
Yawen Han
Shangyao Liu
Jincheng Xu

Plan Overview

Here is the preliminary plan of the team. In order to perform the data analytics of the dataset, works are planned to carry out in the following discovery process:

- Introduction (background overview, object, why it's important, data sources)
- Preliminary data analysis (distribution, investigation, data quality check)
- Data processing (check missing values, outliers, dependency)
- Test and train dataset (30% test, 70% train)
- Modeling (models, metrics, evaluation)
- Results (interpretation/findings, tables, visualizations, test and train error)
- Discussions (conclusion, assumptions, suggestion, limitation, uncertainties, areas of future investigation)
- Reference
- R Shiny Dashboard (list out the questions that we think sales team and marketing team would like to answer and design the dashboard accordingly)

The objective of this project is to explore and analyze the behavior of customers, with the focus of the following sample questions:

- What features influence the behavior to retain customers?
- What can we do to achieve a higher customer retention rate?"

The team will tackle the project from the *Introduction* to have a basic background overview and discussing the significance of the project. Then, the team will perform the *Preliminary Data Analysis* to explore the distribution of each feature and check data quality. Based on the analysis, *Data Processing* will be performed to handling missing values, outliers, correlations and insignificant data values.

Modeling will also be an important part of this project to help the team classify churn customers based on different features provided from the dataset. Our team will explore and seek to identify the most important features that significantly influence the result. Before modeling, the dataset will be split into 30% for testing, and 70% for training. Right now, 3~5 models are planned for modeling. Performance metrics like accuracy, precision will be used to evaluate the performance of each model. After models are chosen, the team will demonstrate the findings and results through tables and visualizations, with further explanations and interpretations of the result to help stakeholders gain better understanding of customer's behavior.

Finally, the team will generate a report and shiny R app for stakeholders to make the data more interpretable. The goal of the app is to allow stakeholders to utilize to gain a more concise and clear understanding of customer behaviors so that they can develop a more relevant programs that will help the company retain customers better.

Progress Summary

Before meeting for Checkpoint 3, our group members have each explored some interesting datasets. And we sat down together in April 1st to briefly introduce our dataset to other group members and finalized the dataset we eventually want to explore for this project. Then we have spent some time to discover this dataset and found interesting problems we want to analyze. Moreover, during the meeting, we have carefully reviewed the rubrics for this project, and looked

Team V (Victory)
Pinmanee Eowpittayakul
Yawen Han
Shangyao Liu
Jincheng Xu

over the deadlines on the upcoming month. Finally, we created a solid timeline to guide us through this project.

Tasks	Timeline	Individual tasks	Meeting time
Determine the dataset and overall goals of the project	04/01/2019	Each person explores interesting dataset on their own and briefly introduce their datasets; finish checkpoint 2; assign tasks for next meeting (explore dataset and come up with cleaning ideas individually)	04/01/2019 2:30pm
Checkpoint 2	04/04/2019		
Explore and clean the dataset	04/08/2019	Discuss cleaning process and finish cleaning dataset; come up with ideas for models and assign the model to each team member	04/08/2019 2:30pm
Discuss models		Each person in charge of developing one model and explain their model to other team members; All team members bring up contributing advices on others' models for further improvement.	04/12/2019 11:00am
Finish and finalize models	04/15/2019	Finalize models; finish checkpoint 3; Assign next task.	04/15/2019 2:30pm
Checkpoint 3	04/18/2019		
Develop R Shiny dashboard	04/23/2019		
Finished report (write-up)	04/28/2019		
Finished report (presentation)	04/30//2019		
Presentation	05/02/2019		

For the timeline table, we have clearly listed the deadlines, and the tasks we need to accomplish for this project. Also, we have set up several meeting times with assigned individual tasks, and topics to discuss over. For now, we have scheduled our meetings until Checkpoint 3, since we want to decide the development of R shiny application after we see our result from the model. In general, our group wants to finish our write-up report by April 28th, and creates our slideshows for the presentation based on the report by April 30th. This timeline table is reasonably designed for each task and is based on the availability of each team member, which gives all of us a clear idea of what to expected for the next meeting and what is left for us to conquer.