Report: Telecom Churn Data Refining

This paper demonstrates the purpose and function of executive data dashboards and the ETL process to produce them. The information was gathered from a telecommunications organization and the industry in general. During the data acquisition process, I have uncovered finer detail than was initially available in some early exploratory data analysis and the subsequent machine learning processes. I will demonstrate through the following that though the raw data points and numerous dimensions were initially intimidating, even with such a relatively small dataset, SQL data analysis and intuitive Tableau dashboards serve to bring order and business intelligence.

The purpose and function of the executive dashboard is to demonstrate who the most profitable customers were, as well as, discover which customers were at highest risk of leaving (churning) the company. As suggested by the Database Marketing Institute, "[a]nnual churn rates for telecommunications companies average between 10 percent and 67 percent" (Hughes, p. 1). The dashboard begins with emphasizing the company's current churn rate (26.5%). The overall key performance indicators for churned and loyal customers are centered on the dashboard. Finally, A national map of monthly revenue allows executives and data analytics peers to compare metrics across states and a category menu allows them to drill down into the dataset.

The additional dataset provides a basis for comparison with publicly available data at Kaggle.com. This dataset was used in a churn prediction project and gave us an opportunity to compare our company's dataset with another company's. The original source of the dataset is from the IBM Samples Team and represents a "telco company that provided home phone and Internet services to 7043 customers in California Q3" (Samples Team, p. 1).

Awareness of the trends in subscriber growth will inject impetus into the executive for expansion and aggressive marketing in our current market as well as encourage investment in competitive markets. It appears "customer churn is particularly problematic," as pointed out by Rohit Chowdary on LinkedIn (Chowdary, p. 1). Indeed, "[r]oughly 75% of the subscribers signing up every year come from another network they are already churners" (Chowdary, p. 1). That customers migrate to and fro in the telecom industry looking is a double-edge sword for our company. However, as our company's churn rate is on the low side of the average annual churn rates, there appears to be a great opportunity to attract new customers, given cost-effective promotion and marketing.

The business intelligence tool used was Tableau. Tableau is an obvious industry leader when it comes to effective and efficient dashboards. Also, as cited at Evolytics, "Tableau's built-in data connections and preparation tools enable analysts to quickly get data in a usable format" (Evolytics, p. 1). The speed with which something may become a useful tool to recognize significant patterns is much greater when compared to statistical analysis using Python or R. The drag and drop interface allows a user to skip the trial and error of writing code and jump straight to the insights, as are demonstrated in the included executive and international dashboards.

Steps used to clean the data include:

- Imported dataset to Python dataframe.
- Renamed columns/variables of survey to easily recognizable features (ex: "Item1" to "TimelyResponse").
- Viewed a description of dataframe, structure (columns & rows) & data types.
- Viewed summary statistics.
- Checked for records with missing data & impute missing data with meaningful measures of central tendency (mean, median or mode) or simply remove outliers that are several standard deviations above the mean.
- Viewed univariate & bivariate visualizations.
- The prepared dataset was extracted & provided as "churn_prepared.csv".
- The tables were sliced out of the "churn_prepared.csv" dataset to resemble the given WGU pgadmin database.
- Referential integrity was enforced by "build[ing] and maintain[ing] logical relationships between tables to avoid logical corruption of data" (w3resource, p. 1). Corresponding

primary key and foreign key (customer_id, contract_id, job_id, location_id and payment_id) relationships were set up to avoid any corruption of the data in SQL joins.

- Finally, the tables that were sliced out of the initial dataset were re-integrated as separate worksheets in one MSExcel csv and saved as “churn_clean_sliced.csv”, which was attached to this submission along with the external IBM telco customer churns dataset.

The dashboard was created in five main steps. First, in bold at the top left of the dashboard, a key metric was set up as the ratio of those customers who churned in the last month relative to the number of customers we began the month with. Second, a table of key performance indicators was placed right at eye level, in a sense, to give executives the ability to compare the features of those loyal customers with respect to those customers who left the company. Third, an color-based, color-blind sensitive, interactive menu was set up at the top right corner to give users of the dashboard the ability to compare and contrast gender with churn decisions. This menu interacts with the geographical layout of customers & revenue, which follow. The fourth step was to design and create a geographical plotting of revenue density by state at the bottom left corner. The fifth and final step was a stacked bar chart showing overlaid revenue from loyal and churned customers. Comparative relationships were demonstrated with the external dataset via the map and barchart visualizations to give executives a sense of a garden-variety competitor in our market.

It is critical that decision-makers & marketers understand that there is an inverse relationship between our target variable of Churn & several of our predictor variables. This suggests that as a customer subscribes to more services that the company provided, an additional port modem or online backup for example, they are less likely to leave the company. Clearly, it is in the best interest of retaining customers to provide them with more services & improve their experience with the company by helping customers understand all the services that are available to them as a subscriber, not simple mobile phone service. Given the negative coefficients of additional services, we suggest additional marketing efforts for contracts & internet services as those with a contract appear less likely to leave the company.

Also, with such a direct linear relationship between bandwidth used yearly & tenure with the telecom company it makes sense to suggest the company do everything within marketing & customer service capability to retain the customers gained as the longer they stay with the company the more bandwidth they tend to use. This would include making sure that fixes to customer problems are prompt & that the equipment provided is high quality to avoid fewer replacements of equipment.

Limitations given the telecom company dataset (‘churn_clean.csv’ provided by WGU) are that the data are not coming from a real world warehouse. In this scenario, it is as though I initiated and gathered the data. So, I am not able to reach out to the staff that organized & gathered this information to ask them why certain NAs are there, why are fields such as age or yearly bandwidth usage missing information that might be relevant to answering questions about customer retention or churn. In a real world project, you would be able to go down to the department where these folks worked and fill in the empty fields or discover why fields are left blank. The dataset acquired from Kaggle and IBM (‘WA_Fn-UseC_-Telco-Customer-Churn.csv’) also has the same limitations.

It should be pointed out, as well, that it is a great limitation that this process is happening rather in a vacuum. That the process of acquiring, cleaning, exploring, analyzing and presenting this data was done solely by me in a small office would seem to remove all credibility in possible recommendations for action. That is not to say that my skills are in question as a young analyst. It is only admitting the short-coming of a young, single, white male doing the project. Perhaps the project could provide much greater benefit with a larger team of analysts working on this project. A focused team of more experienced data scientists as well as culturally-diverse individuals would undoubtedly come up with greater and deeper insights for the sake of churn prediction and meaningful business intelligence.

D. Web Sources

- Kaggle. (2021). Telecom Churn Prediction. www.kaggle.com. https://www.kaggle.com/bandiatindra/telecom-churn-prediction/data.
- Samples Team. (2019, July 11). Telco customer churn (11.1.3+). www.community.ibm.com. https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113

E. Sources

- Chowdary, R. (2021, May 26). Impact of churn in telecom Industry. How BI can potentially solve the problem? LinkedIn. https://www.linkedin.com/pulse/impact-churn-telecom-industry-how-bi-can-potentially-solve-chowdary/
- Evolytics. (2021, September 23). Top 5 Reasons Your Company Needs Tableau. Evolytics. https://evolytics.com/blog/5-reasons-your-company-needs-tableau/
- Hughes, A. (2021, September 23). Churn reduction in the telecom industry. DB Marketing. Database Marketing Institute. http://www.dbmarketing.com/telecom/churnreduction.html
- W3resource. (2021, September 23). SQL joining through referential integrity. www.w3resource.com. https://www.w3resource.com/sql/joins/joining-tables-through-referential-integrity.php