31.12.2019

Deliverable 2 Project Report

Project Title: Analytics of Telco Customer churn data for customer retention under an optimized prediction model.

Problem:

- ★ Telco is having a difficult time at making the necessary adjustments to the company in order to retain their customers. In order to analyze and provide a solution to this problem as our team, we divided our solution to the problem into sub-solutions such as optimization of our data analysis models and statistical inferences we derived from these models where our driving ambition was to predict the customers who were likely to churn in the following days/months.
- ★ Specifically, our team focused on specific attributes for the solution of the project such as the columns of “Churn”, “Tenure”, “Monthly Charges”. The dataset is going to be more accurately and elegantly described in the following sections.

Executive Summary:

First of all; in terms of our analysis and observations which we built our analysis and inference upon, we utilized the Telco attributes (the most relevant ones which we determined in the modeling tryouts later on) as much as we could in order to decrease the amount of canceled subscriptions. Therefore, our analysis heavily relied on the “Tenure” information that Telco has accumulated under its database. After fixing and replacing for the missing values on the dataset attributes such as filling in the null values in the columns with the mean (the most frequent value for string types); we leaned our analysis towards the descriptive part. We specifically tested for the relationship between the loyalty of customers and how much they were charged by Telco on a regular basis. For this stage, different shuffles of the relevant attributes (charges, tenure, churn, senior citizen) were plotted using both Tableau and KNIME.

On the other hand, while conducting these operations, we also managed to focus on the business perspective of mining, therefore, we plotted different variations of ROC and Lift charts for the selected attributes provided above. Our examinations concluded that there was/has been significant positive correlation between the monthly charges and tenure rates, signaling that the most loyal customers were paying on a regular basis and will continue on that alignment if there wasn't/won't be a quality change in terms of Telco's customer service. Next, after determining the selected attributes that we utilized in our models, we calculated for some performance metrics such as recall, precision, and F-1 measure. Since our determination was to obtain the best results we could obtain for recall and precision along with a low misclassification rate using different classifiers such as Naïve Bayes, decision tree and so forth, our results turned out to be quite promising. Reflecting on our best scores, our best accuracy rating was around 80% and our F-1 measure consisting of a combination of recall and precision metrics revealed that we were successful enough to detect customers who churn out of the true customer churn labels in the dataset although we were not able to obtain a high precision score(around 0.5). Overall, we were satisfied with our inferences and our models provided "optimal" results although with different parameter optimization coverage of attributes, model optimization may be improved for future usage.

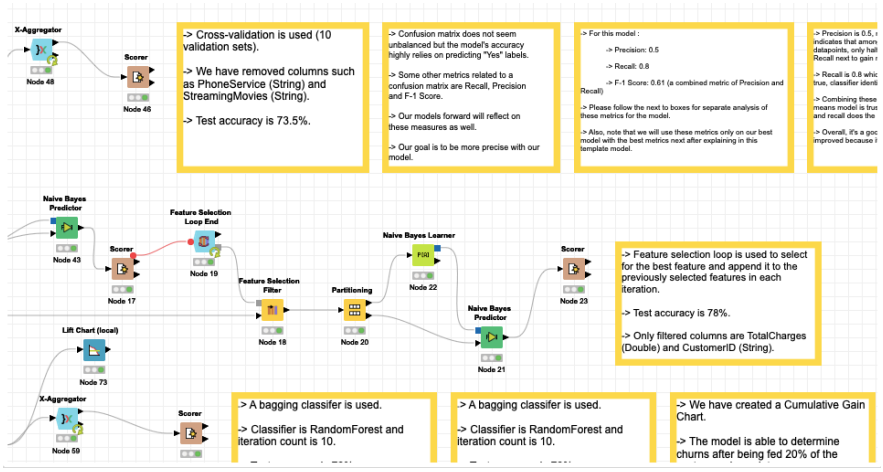

Figure 1. Example KNIME Workflow – Models & Inferences & Results

1. Data Understanding

Describing Data

During this phase of the project, we evaluated what type of data we were dealing with and how could we represent the true value of the dataset in our project.

- ★ Our dataset consisted both categorical and numerical data types. Categorical category was contained of binary (yes-no) as observable in columns churn and types of services included in the dataset.
- ★ Numerical types were split into integer and ratio values. Ratio values include absolute zero points. In the dataset, ratio values referred to the charges made both monthly and total. On the other hand, integer values represented the length of the subscription for a customer as seen in the tenure column.

Note that since the dataset mostly consisted columns of nominal types, one would guess that these attributes would have more weight in the model implementation. However, in our research we concluded that although they have a very considerable effect, the model performance attains its peak when combined with the right integer and ratio columns; tenure, monthly charges and total charges.

Verifying Data Quality

Our dataset was clean, which means we did not have to fill in missing values or do anything else for model preparation. What we did was more of a filtering job rather than a cleaning job. We filtered out the columns that caused unbalanced confusion matrices (low recall, precision and F-1 scores) and low misclassification rates. A good example is CustomerID column.

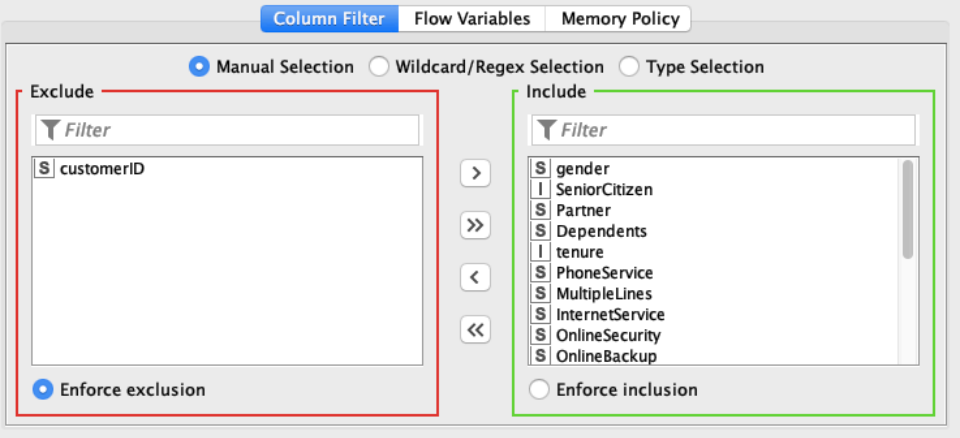

Figure 2. Example Filter - CustomerID column removal (KNIME)

2. Data Preparation

Selecting Data

★ Since we designed our prediction model for Telco to maximize customer retention, the selected attributes had to bind well into the model's parameter optimization and target value, churn(yes-no) in order to obtain a balanced and accurate confusion matrix result.

★ Even though we designed several models and performed several filter operations, the models which we used the most were monthly charges, total charges, tenure, all kinds of service columns with random shuffling while transitioning from one model to another and the target value churn. The model performed the best under these attributes.

★ Additionally, we performed parameter optimization for most attributes.

★ The best response was from the logits function, a probabilistic activation function for our data in the cost sensitive classifier that blended well with the predetermined columns, yielding an accuracy of 80% with 0.62 F-measure which is acceptable.

3. Modeling

Selecting Modeling Techniques

★ Our dataset was clean going onto this phase of the project, therefore the data that fed our models was the best version of the dataset that we could utilize for our mining.

★ We implemented exactly 6 models of KNIME with different classifiers, with different activation functions, with different partitioning. In this part, we are going to centralize on the top two that we implemented.

1. Bagging with Random Forest

In this model, we filtered out the CustomerID because this column no matter the classifier, only serves as a corrupted data column which has no relation to any datapoint from another datapoint in that column. After filtering out the CustomerID column, we split the data in into two partitions 66.6% training and 33.3%testing. Then, we cross validated using KNIME X-partitioner node and obtained an accuracy of 79% in the bagging classifier where the bagging iteration count was 10 and the model ran on a single thread, execution slot. The parameter optimization node was unnecessary since we could control for the iteration with the classifier itself. As a final remark, we created a Lift Chart where it's obvious that our model performs as intended after getting fed with 20% of the data.

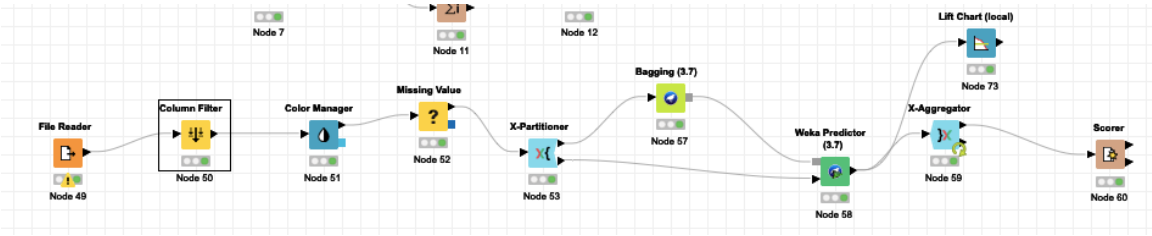

Figure 3. Bagging Classifier Workflow – 79% Accuracy, 20% and above catching threshold with Lift Chart (KNIME)

2. Cost Sensitive Classifier with Logits Function

After bagging, we realized that we were coming to a sense of understanding what the most relevant nominal data we could work around with for our best model. The resulting columns of our discussion were tenure, monthly charges, total charges and senior

citizenship. Hence, we could give equal shots to different types of data in our model with a logits function. Logits function normalizes to data to either between 0 & 1 or –1 & 1 so distinct values do not interfere with the model's performance. Also, we performed parameter optimization by brute searching within an interval of 100 iterations. Finally, we plotted a ROC curve with respect to monthly charges column where above 20% under charges attribute, the model performance was very high, the area under the ROC curve was 0.7.

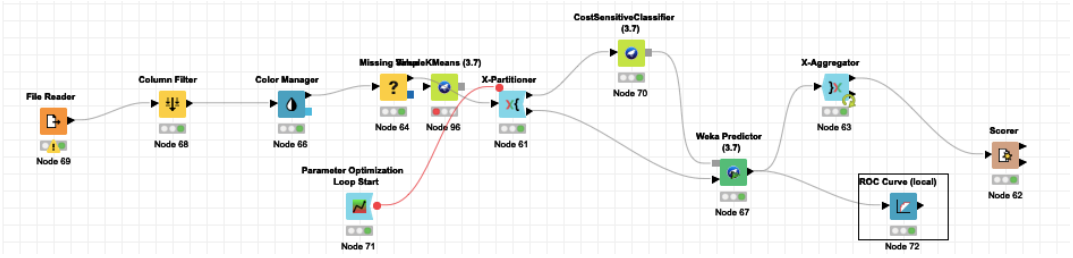

Figure 4. Cost Sensitive Classifier Workflow – 80% Accuracy, 0.7 ROC Area under Monthly Charges (KNIME)

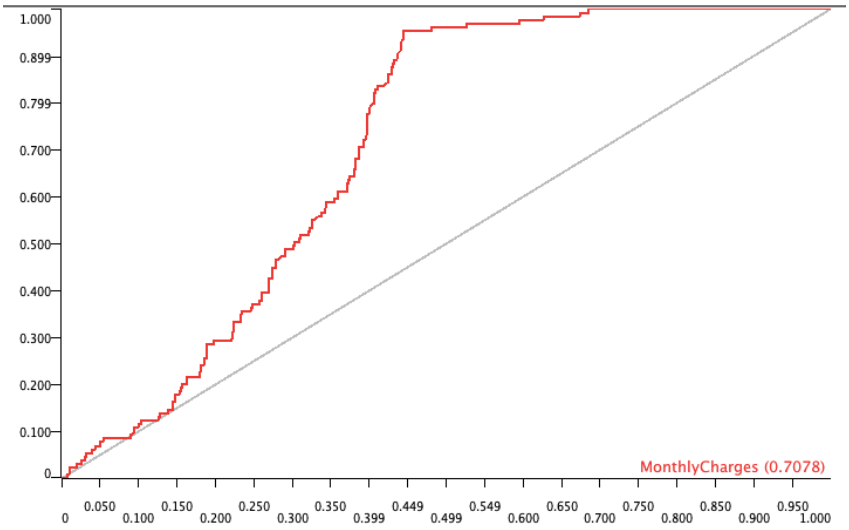

Figure 5. ROC Curve, referring to MODEL 2 (KNIME)

4. Conclusion

★ The below chart that we plotted in the first place presented that Telco was losing customers and precautions needed to be taken.

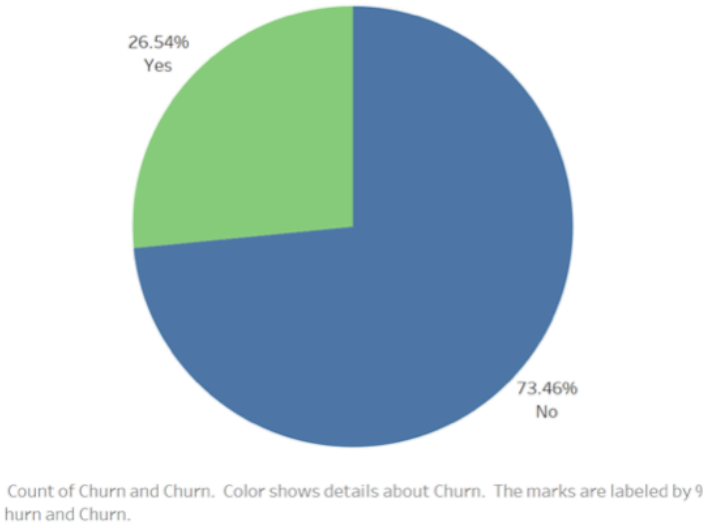

Figure 6. Churn Statistics (TABLEAU)

★ Gender column was filtered out in some algorithms because Tableau presented that allocation of churn rates between genders was not feasible for a solution.

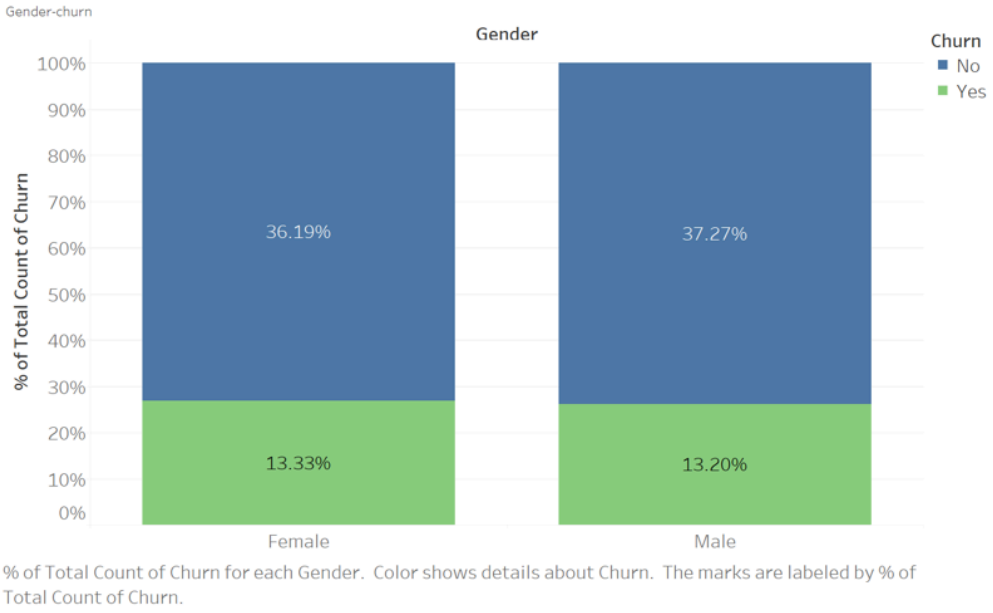

Figure 7. Gender vs Churn (TABLEAU)

- ★ In the end of our mining efforts, as explained prior to this section in the report, we implemented several models with several algorithms and created the relevant charts and plots using KNIME and Tableau.
- ★ Our churn analysis was successful on both descriptive, inferential and predictive standards.
    - We matched he prediction standards because we obtained an unproblematic confusion matrix with a high accuracy score of 80% without overfit.
    - We matched the inferential standards because statistically we were able to utilize performance metrics such as precision, recall and F-measure each above 0.5. Furthermore, we used these metrics to choose the best algorithms for our workflows in KNIME.
    - We matched the descriptive standards because with Tableau, we were able to generate insightful plots where we determined the attributes contributing to or diminishing churn rates.
- ★ Tenure dominates monthly charges above high intervals, and this allows the loyal customers to stick with the company. On the other hand, charging the newcomers with a high subscription fee results in an increase in churn rates for Telco. Our models and plots have identified this problem, and this could be fixed by delaying the extra charges on newcomers or offering them promotions. Loyal customer overcharge could go on and is very beneficial. Although there are occasional churns, expected value is well towards not churning (see Figure 6).

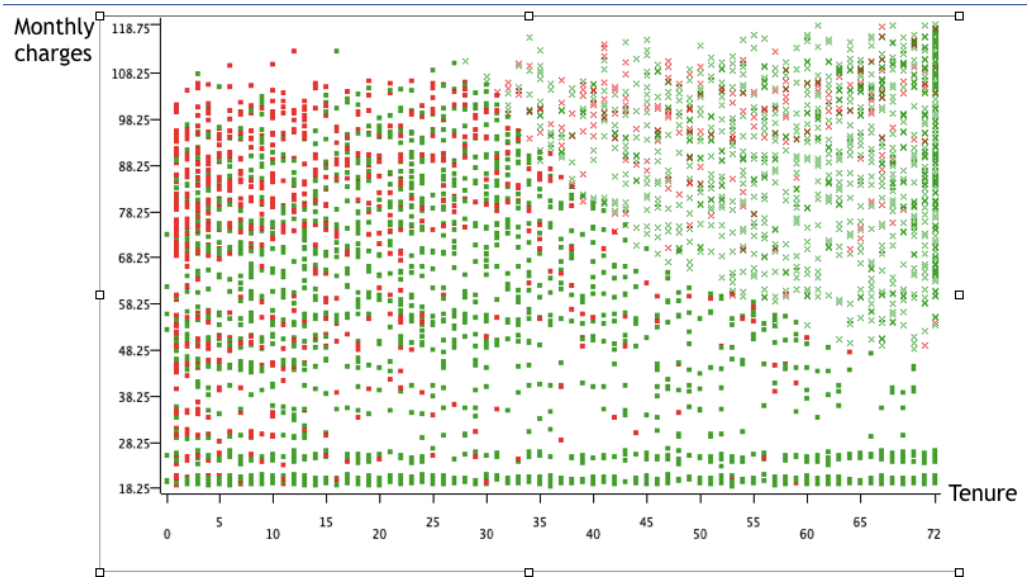

Figure 8. Monthly Charges vs Tenure on Churn Rates (Green represents "Not Churn") (TABLEAU)

★ Longer the contract span gets, less likely the customers are going to churn. Therefore, offering long term contracts with small discounts but overall expected increasing marginal returns may be a sensible solution for Telco as well.

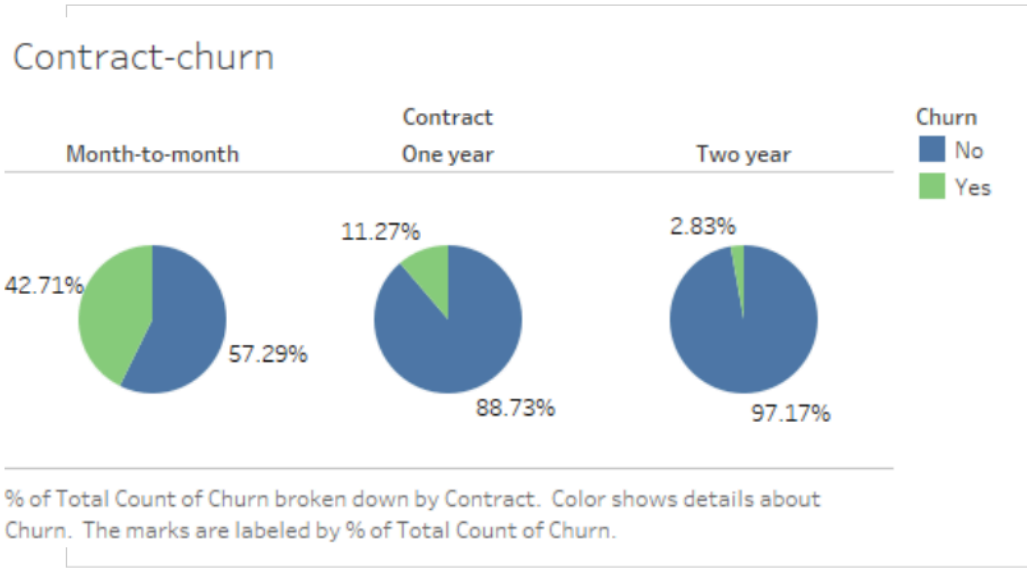

Figure 9. Contract Span vs Churn (TABLEAU)

✗ ***Important Remark:*** *We plotted several other graphs in order to construct for our models but including all of them in this document would exceed our page limitation. Therefore, we are providing them in the .zip extension which is our submission file. The report concentrates on every key aspect and most of the minor coverage. However, please refer to the remaining files in the .zip for total coverage.*