Customer Churn Prediction for Telco Companies

Kuluhan Binici T07902133 **Berke Uğurlu T07901110**

Jack Walkinshaw T07602101 **Jack Lin D06546002**

1. Introduction

The dataset was released in 2015 IBM website which is used as an example by Watsons Analytics to understand why customers leave (churn) from telecoms who provide DSL services to their cable competitors. There are differences between DSL internet and cable internet in terms of technology and cost. DSL services are based on the standard copper phone lines which most people have at home before the popularity of mobile phones. DSL also has dedicated bandwidth and cheaper whereas cable internet is faster, more expensive, and is based on cable TV.

Although the dataset was first released to analyze why customers moved from DSL to cable, it is unavoidable to consider the rise of LTE services for the customer churn. Since TeliaSonera launched the world's first LTE services in 2009, the speed of wireless network has increased from Cat.3 of 100 Mbit/s to Cat.11 of 600 Mbits/s where the standard speed of DSL ranges between 2 Mbit/s to 100 Mbits/s and cable ranges from 10 Mbits to 2,000 Mbits/s. As a result, telecoms have tried to increase the revenue for their landline business by separating phone services from internet services, promoting fibre to home, and providing streaming TV/movie services.

Some variables in the data contain redundant information. For examples, “MultipleLines” variable includes the service availability of phone services in “PhoneService” and similarly, streamingTV shows if “InternetService” is available. However, the choice of services does not indicate the costs for the service. Furthermore, there are many factors that could be resulting in people leaving the company or the contracts such as the services they are using, the costs and times of payment, the customer demographic and how long the customer has been at the company are all features which the company. It is extremely useful for the company to know who is likely to leave initially and what factors are the key drivers for this so they can target that customer demographic initially and maintain retention of customers. Especially due to things like telecommunication companies that rely heavily on customers who have a high tenure as they are the ones who are unlikely to look for better deals and often mean more money for the company.

The current data has 7044 customers and is sufficient to determine customer churn as we have a multitude of variables 21 variables which includes data on Customer Demographic, the services customers have, account information and history with the company and information of if they have churned or not which means we can create learning and testing dataset and then use this algorithm we create to them be applied to a larger dataset of the current customers upon completion . We create the percentage for how likely they are to leave so that the company can address those who are most at risk first then move to the less at risk.

2. Treatments and Algorithms

2.1 Imbalanced data - Random under Sampling

Due to the imbalance in data of 1:3.6 the algorithm was distinguishing negative samples better than positive ones which resulted in high type 2 error (false positives). So, we chose Random Under Sampling to balance the data as we had sufficient amount of data and the ratio difference was not extreme. Thus, the risk of overfitting was mitigated by the benefit it brought to the accuracy.

2.2 Erroneous data and treatments due to Random Under Sampling

Random Undersampling also caused erroneous column data type classification. Thus, reclassification of string numeric columns to float and categorical from float to string. Empty entries in 'Total Charge' column was replaced by 0 as these customers had a tenure of 0 which meant they had not yet paid the company for their services yet. Due to the data loss from Random Undersampling we chose a train test split of 80/20 in order to allow the model to learn effectively and have sufficient amount of data.

2.3 Converting categorical data to numeric: One-Hot encoding

One-Hot encoding gives integer representations of the categorical data. We chose not to do the Label Encoding as the model assumed an order between the variables the model represents them in a Euclidean space. This order does not exist, so we chose One-Hot encoding as it is nor influenced by this issue. We chose to create a n element array which corresponds to n-1 new columns in order to mean one less column compared to the n=n One-Hot Encoding method which results in a redundant column for each variable encoded.

2.4 Selecting features and labels:

CustomerID was removed as it served no meaning other than once the data has been trained. Additionally, churn was set as labels as the model is supervised dependant on Churn.

3. Learning Models

3.1 SVM with kernel function

We picked SVM as one of our classification algorithms based on the knowledge that it tries to maximize the confidence in predictions while outputting a linear hyperplane that separates data belonging to different classes from each other. Since the data is in most cases not linearly separable, we also used different kernel functions to map our non-linearly separable data to a different space in order to make it linearly separable. The kernels we used were; rbf, which is a generalized version of gaussian kernel, polynomial and sigmoid. We picked the best performing kernel, which is rbf based on the 3 metrics we picked for evaluation, among these to compare with the gradient boosting tree.

3.2 Gradient Boosting Tree

When picking our other classifier to be gradient boosting tree, we considered that it is a model that is capable of handling both regression and classification tasks, supports both numerical and

categorical data as inputs and outputs strong learners by combining several weak decision tree learners.

4. Evaluating and Benchmarking the Results

For evaluating our results, we decided to consider 3 different evaluation metrics that are commonly used in machine learning, which are:

4.1 Accuracy

We considered accuracy in order to see how many testing data samples our models classified correctly. The results indicated that rbf kernel was the best performing in terms of this metric among other kernels. In the comparison between boosting tree and SVM, even though boosting tree performed slightly better, we cannot say that one model outperformed the other.

4.2 Confusion matrix

Confusion matrix let us know how well our models learned to distinguish positive and negative samples. By the TP, FP, FN, TN values it contains, we can also calculate the metrics like Recall, Precision, Specificity, Accuracy. We became aware of the problem that the imbalance between the numbers of positive and negative samples caused by looking at this metric and decided to apply random under-sampling to circumvent the problem. Based on this metric rbf kernel was again the best performing among all kernels and in its comparison with the boosting tree there again wasn't a clear winner even though boosting tree performed slightly better.

4.3 ROC curve

ROC curve also is a metric to monitor how well our classifier does the separation between classes and it is highly related to the confusion matrix. It also allowed us to calculate Area Under the Curve (AUC) value, which is a measure of, again, how well the separation is made. The bigger the AUC value is, the better the separation is. In our case rbf was the best performing also based on this metric and again it was slightly worse performing than boosting tree.

5. Interpretations: Directional Feature Contributions (DFC's)

We used DFCs to evaluate features of our dataset based on our model. We used DFC because It shows what features were used most to make a prediction by the algorithm. Due to random-under-sampling and the dataset separation by 80/20 ratio. We got 747 customers remaining. Each customer gave us different DFC based on the algorithm we used. In our most of the DFCs we saw that tenure is the most important feature for our model. And gender feature decreased the accuracy of our model.

5.1 Gain based feature importance

We used Gain-based feature importance and we saw the effects of all the features we have. We sum all the DFCs we have. In our model, we saw that tenure is the most important feature for our gradient boosting tree algorithm. And gender feature decreased the accuracy of our model.

5.2 Average Absolute DFC's

Average absolute DFC takes absolute values of all the DFCs for each person and averages the feature importance together. Average absolute DFC's are consistent with Gain-based feature Importance. We concluded that tenure and gender are the most important features for our model.

6. Conclusion

We evaluate our features based on the DFCs. We saw that Tenure increases the accuracy of the algorithm; however, gender decreases the accuracy of the algorithm. Based on this result, we can say that most of the accuracy came from tenure. Model error was due to gender, the other features appear to play a minor role in our algorithm. We compared Gain-based feature importance with Average absolute DFCs. We saw that our Average absolute DFC's have the same order with Gain-based feature Importance. An area for further investigation should be configuring the model to have tenure as the label in order to determine which factors lead to a customer having a high tenure.

Tenure was highly correlated to churn according to our DFC's. It is possible that this was due to the One-Hot Encoding method as tenure had a much greater variety of values within its column compared to all other categorical values. This resulted in many more columns influencing tenure than any other variable. Thus, it is believed that this may have influenced the weighting of the model to favour this variable above all else. This suggests a possibility of further analysis in finding an alternative to both One-Hot Encoding and the Label Encoding method for tenure to test this hypothesis.

Several of the variables were highly correlated, such as MonthlyCharges and TotalCharges. This might not be significantly beneficial compared to considering them as one exclusively. Since the decision tree looks at the information gain, in theory the prediction from the information gain should not be too different highly correlated variables. However, it may be worth investigating by combining highly correlated variables to see if it increases information gain and the overall influence to churn.

In real world, we can say that tenure shows the confidence between customers and companies. If a customer has stayed in the company for many years, that customer will very likely stay in the company. We saw the same result in model. Our outputs on the model contradicts our intuitions as there were so many insignificant features. For instance, monthly charge appeared to be insignificant to churn based off the DFC's. This information would be highly relevant for the management of the business in ensuring they focus on areas most likely to have the highest return on investment rather than their intuitions. For example, one insight might be that the consumers are likely to choose a contract that is proportional to their income and the types of entertainment that they require and thus is unlikely to be a factor of their churn.