Team 4

**Telecom Customer Churn Data Set**

**Executive Summary**

The key business goal for Telecom Customer Churn dataset is to analyze various customer behaviors, identify the plausible correlations between the variables and develop a rational plan for customer retention. The data set was downloaded from Watson Analytics and comprised of 21 variables with 7043 records. Based on the available customer information such as monthly rates, customer plans, value added offerings, monthly rentals', total charges, current customer churning, etc., we developed the best model that can predict the prospective customer churn.

**Data Processing and Cleaning**

The data set in generally appeared as clean but the following steps were done to make it ready for effective modeling.

**Excluded the missing rows for ‘Total Charges’ column**: We excluded the 11 rows because there were ample analogous data combinations where customer is not churning out.

**Recoding the Contract:** We did not see much difference between the patterns of one year and two year contracts. So recoded ‘Two year’ and ‘One Year’ to ‘LongTerm’ and ‘month to month’ to ‘ShortTerm’ for more scalability while analyzing.

**Multivariate:** Conducted multivariate on ‘Tenure’, ‘Monthly Charges’, and ‘Total Charges’ and identified a positive correlation between them and hence noted this observation while selecting the variables during modeling.

**Data Type Conversion:** Evaluated the significance of each variable data types. Changed the data type of senior Citizen column to nominal.

**Outlier Analysis:** Outlier analysis using Mahalonobis distances was checked for ‘Tenure’, ‘Monthly Charges’, and ‘Total Charges’ and didn’t observe any outliers.

**PCA:** Created Principal Components using ‘Tenure’, ‘Monthly Charges’, and ‘Total Charges’ for enhancing the various model, detailed in the next section. We observed that the PCA is not adding any values to the model. So removed the column.

**Distributions and ‘Fit Y/X’:** Ran distribution on each variable and analyzed the proportions of each variable in the data set. Analyzed each variable against the churn individually and identified the individual skew of each variable against ‘Churn’.

**Variable Selection:** On our primary analysis, ‘MultipleLines’ is a good representative of “Dependents’ and ‘Partner’. So these 2 variables are dropped for modeling. Excluded the variables ‘OnlineSecurity’, ‘OnlineBackup’, ‘DeviceProtection’, ‘StreamingTV’ and ‘StreamingMovies’ because these variables were correlated and dependent to ‘InternetService’. Based on the experimentations done with all the variables, other than above 7 values, and by considering P-Value and multiple modeling outputs, refer Table A, we had concluded the predictors as ‘Tenure’, ‘MultipleLines’, ‘Internet Service’ and ‘Contract’.

**Cleaning:** We had identified 2 exceptions while trying to improve the model.

**1.** Excluded 214 rows where contract is LongTerm and Churn is Yes. These 214 rows are about 6% among the LongTerm churn data. We are ignoring this 6% because there is no correlation with any other variable that may be causing this churning and the churning looks purely random and misguiding to the model. We also observed that the predicted value for these records were false. So eliminating them from training the model seems justified and more effective.

2. Excluded 160 rows when Tenure is <=6, Churn is No and Internet Service is Fiber Optic. The common behavior of low tenured customers when they use expensive Fiber Optic Internet connection is high churn turn out. We had identified oddly behaving 160 rows. Thus we are removing these rational outliers by assuming that these non-churning entries will not impact the model goal.

The above 2 exclusions are improving the model accuracy and misclassification rate. Please refer **Table A** for more details.

**Training/Validation Ratio**: Created the new column ‘Validation’ by splitting the ‘Training’ to ‘Validation’ in the ratio 70:30.

**Modeling**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table A (Model Experiments and Observations) | | | | | | | | | | |
| Ver. | Model | Variables | R^2 Train | R^2 Valid. | Misc. % | MR Type1 | MR Type2 | ROC | Comments |
| 1 | Decision tree | 12(G,SC,P,D,PS,ML,IS,TS,C,PB,PM,TC) | 27.00% | 24.70% | 20.30% | 15.60% | 36.70% | 83.50% | Number of splits -12 |
| 2 | Logistic | 12(G,SC,P,D,PS,ML,IS,TS,C,PB,PM,TC) | 25.60% | 26.02% | 20.26% | 16.18% | 35.45% | 83.30% |  |
| 3 | Neural | 12(G,SC,P,D,PS,ML,IS,TS,C,PB,PM,TC) | 27.20% | 26.90% | 19.90% | 16.60% | 33.57% | 84.10% | TanH(3,0) |
| 4 | Decision tree | 7(G,SC,PS,IS,TS,C,TC) | 28.80% | 26.70% | 19.90% | 16.06% | 34.47% | 84.50% | Number of splits -36 |
| 5 | Logistic | 7(G,SC,PS,IS,TS,C,TC) | 24.10% | 25.40% | 21.40% | 16.95% | 38.33% | 82.20% |  |
| 6 | Neural | 7(G,SC,PS,IS,TS,C,TC) | 25.10% | 26.50% | 21.06% | 17.25% | 36.51% | 83.05% | TanH(3,0) |
| 7 | Logistic | 5(T,C,IS,TC,D) | 25.62% | 25.53% | 21.13% | 17.37% | 36.54% | 83.41% |  |
| 8 | Neural | 5(T,C,IS,TC,D) | 26.75% | 26.73% | 20.88% | 16.28% | 37.36% | 83.79% | TanH(3,0) |
| 9 | Decision tree | 5(T,C,IS,TC,D) | 26.50% | 26.50% | 21.05% | 16.88% | 37.03% | 83.33% | Number of splits -12 |
| 10 | Logistic | 4(T,C,IS,G) | 25.01% | 25.69% | 21.45% | 17.40% | 37.63% | 83.13% |  |
| 11 | Neural | 4(T,C,IS,G) | 26.01% | 26.42% | 21.53% | 17.94% | 37.04% | 83.45% | TanH(3,0) |
| 12 | Decision tree | 4(T,C,IS,G) | 26.55% | 26.47% | 21.05% | 16.88% | 37.03% | 83.33% | Number of splits -12 |
| 13 | Decision tree | 5(T,C,IS,ML,TC) | 27.69% | 26.71% | 20.89% | 12.67% | 40.61% | 84.05% | Number of splits -16 |
| 14 | Logistic | 4(T,C,IS,ML) | 25.72% | 25.87% | 20.82% | 16.46% | 36.92% | 83.42% |  |
| 15 | Neural | 4(T,C,IS,ML) | 27.44% | 26.87% | 20.49% | 16.65% | 35.52% | 84.23% | TanH(3,0) |
| 16 | Decision tree | 4(T,C,IS,ML) | 27.69% | 26.71% | 20.89% | 12.67% | 40.61% | 84.05% | Number of splits -16 |
| 17 | Decision tree | 4(T,C,IS,ML) | 40.46% | 40.40% | 18.45% | 8.81% | 40.61% | 88.97% | Excluding Exception 1 |
| 18 | Decision tree | 4(T,C,IS,ML) | 35.57% | 33.76% | 18.61% | 13.67% | 33.25% | 86.56% | Excluding Exception 2 |
| 19 | Decision tree | 4(T,C,IS,ML) | 48.89% | 48.26% | 16.04% | 9.98% | 33.25% | 91.51% | Excluding both exceptions |

**Abbreviations**: Gender: G, SeniorCitizen: SC, Partner: P, Dependents: D, Tenure: T, PhoneService: PS, MultipleLines: ML, InternetService: IS, OnlineSecurity: OS, OnlineBackup: OB, DeviceProtection: DP, TechSupport: TS, StreamingTV: STV, StreamingMovies: SM, Contract: Co, PaperlessBilling: PB, PaymentMethod: PM, MonthlyCharges: MC, TotalCharges: TC

**The best model that we chose from the above**

Nineteen different modeling combinations were conducted during the project tenure by constantly considering the business value addition in telecom customer churn. We selected the simple Decision Tree model with the lesser number of variables without compromising the efficiency of the predicting formula. This model is having the least Misclassification Rate, competitive training/validation R^2, lowest Type1 Error and the best ROC.

**Conclusion**

The prediction model and data analysis shows below trends of customer churning:

* More than 90% of the long term customers tend to stay with the company
* 40% of short term fiber optics user tend to leave the company
* People using no internet services stay with the company majority of the time

**The additional variables that we think that could add value to this model**

The challenges that we faced were lack of some kind of information like Age Group, Network Coverage, Customer Income, Call Duration, Call Drop, etc. We believe that these factors could be some of the key insights to determine customer churn out behavior in the telecom industry. We made a lot of assumptions and transformed some factors and rows, to make our model more meaningful.

**Recommendations for the Company**

* The company should aggressively focus on customers whom are staying for low tenure, identify their problems and come up with plans to prolong their subscription.
* Company should analyze the problems encountered by customers using Internet Services because the customers with no Internet Services are tending to stay with the company.
* Company should strategize plans for Fiber Optics users whom are churning out, especially in the early period [mostly tenure less than 1 year], by evaluating their problems such as expensive data plans, network coverage and connectivity issues etc. and steps should be taken to solve them.
* Identify the shortcomings of Short Term customer contracts and improve the deficiencies. On the other hand, continue the services provided in the Long Term contracts to retain them.