Telco Customer Churn Analysis

Robert Carlton
GTEID XXXXXX982
Georgia Institute of Technology

***Abstract*—Abstract: Customer churn can be a difficult problem for many businesses, especially those is competitive markets such as telecommunications. Telcos can reduce customer turnover if they can proactively identify those customers likely to churn. In this analysis, I examine regression, classification and tree-based models to determine which most accurately can predict churn. In examining the telco customer churn data set, I found that there are statistically significant differences between mean tenure, monthly charges and total charges for customer who churn versus those who do not, potentially leaving significant revenue untapped due to customer churn. In evaluating logistical regression, k-nearest neighbors and several tree based methods, I found telcos have an opportunity to reduce churn and extend customer relationships and increase revenues. Random forests provided the highest accuracy of the model tested however other metrics may be equally valuable depending on the telco approach to retaining customers.**

***Keywords*—churn, knn, naive bayes, decision tree, gradient boosting.**

I. Introduction

This analysis will examine and analyze telco customer churn data. Customer churn, alternatively churn-rate, is the annual rate at which customers stop subscribing to a service. To remain viable, businesses must continually acquire new customers for their service and retain existing customer equal to or in excess of their churn rate.

There can be many reasons for customer churn, including:

- changing customer needs - customers may require different features or entirely different services over time that their current service provider may not offer,
- competitive offers - competitors may offer similar services at a lower cost or enhanced services at the same cost that providing an enticement for customers to shift to another service provider,
- service dissatisfaction - customers may become dissatisfied with the company/organization's service offerings, or they may not meet the customer performance needs, encouraging them to consider other service providers.

These are just a few of the reasons customers may churn. Understanding the dynamics of customer churn - which customers are likely to churn, when customer's might churn, products that can enhance customer lifetime, etc., - can open the door to strategies that reduce churn, lengthen customer relationships and increase profitability.

II. Data Source

The data for this analysis was acquired from Kaggle. It is the Telco Customer Churn data set and consists of 7043 customers, each with 21 features/attributes. The data set can be found online at this URL: https://www.kaggle.com/datasets/blastchar/telco-customer-churn?select=WA_Fn-UseC_-Telco-Customer-Churn.csv .

The data set includes demographics features, such as gender and partner status, and product/service features including tenure and types of products/services used by each customer.

Importantly, the data set includes customer tenure, which may help identify when churn intervention strategies would be most appropriate; and a churn identifier which can help with classifying customers with a high likelihood of churn.

Variable Name	Description	Type	Levels
customerID	Unique customer id for each customer record	Integer	NA
gender	Gender of customer	Nominal, categorical	Female, Male
SeniorCitizen	Whether the customer is a senior citizen	Nominal, categorical	No, Yes
Partner	Whether the customer has a partner	Nominal, categorical	No, Yes
Dependents	Whether the customer has dependents	Nominal, categorical	No, Yes
tenure	Customer tenure measured in months	Integer	NA
PhoneService	Whether the customer has phone service	Nominal, categorical	No, Yes
MultipleLines	If the customer has multiple lines	Nominal, categorical	No phone svc, No, Yes
InternetService	Type of internet service	Nominal, categorical	No, DSL, Fiber optic
OnlineSecurity	Type of online security service	Nominal, categorical	No, Yes, No internet Svc.
OnlineBackup	Type of backup service	Nominal, categorical	No, Yes, No internet Svc.
DeviceProtection	Whether the customer has device protection	Nominal, categorical	No, Yes, No internet Svc.
TechSupport	If the customer has a tech support package	Nominal, categorical	No, Yes, No internet Svc
StreamingTV	Whether the customer has streaming tv	Nominal, categorical	No, Yes, No internet Svc
StreamingMovies	Whether the customer has streaming movies	Nominal, categorical	No, Yes, No internet Svc
Contract	If the customer is on a contract, and if so, for what time period	Nominal, categorical	Month-to-month, 1 year, 2 year
PaperlessBilling	Whether the customer is on paperless billing	Nominal, categorical	No, Yes

Variable Name	Description	Type	Levels
PaymentMethod	The type of payment method used by the customer	Nominal, categorical	Bank Transfer, CC, e-Check, Mailed
MonthlyCharges	Recent monthly charges for the customer	Integer	NA
TotalCharges	Total charges for the customer	Integer	NA
Churn	Whether the customer has churned	Nominal, categorical	No, Yes

Table 1- Variable names, descriptions, types and levels for the telco churn data set. Class highlighted in green.

III. Dataset Analysis

To better understand the telco churn data set, several exploratory analyses were conducted; include histograms to understand the distributions of the variables, analysis of missing data and several different approaches to imputing missing data across the independent variables, checking for balanced classification data and correlation analysis of all variables. The results of these analysis and any remedial actions taken to improve the data are briefly described below. (Note: Fully commented code can be found in the appendix for all analyses.)

Data distribution – histograms of all the data were analyzed to better understand how the data is distribution (gaussian, binomial, etc.) Generally, the data was not distributed normally for the variables. Many of the categorical independent variables followed a Bernoulli distribution having two, or three, possible outcomes. MonthlyCharges appeared to follow a uniform distribution, while tenure appeared to follow a Beta distribution. None of the data appeared to require special handling.

Violin plots of the numeric data fields (e.g. tenure, MonthlyCharges and TotalCharge) were develop to provide insight into differences in how the data were distributed for churners and non-churners.

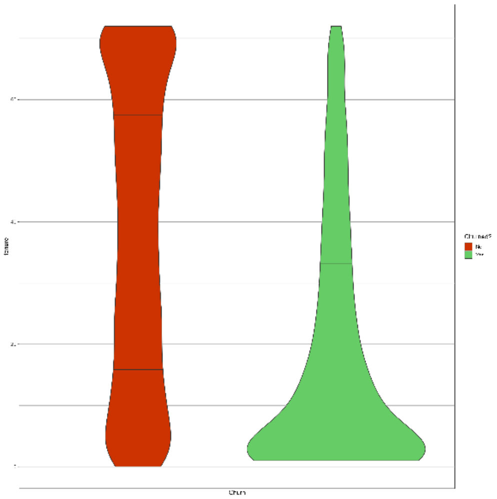


Figure 1 - Violin diagram of Tenure. Churned customers are shown in green.

In Figure 1, the violin plot of tenure by churners vs non-churners, we see that tenure for non-churners (in red) appears to be uniformly distributed across the customer lifecycle whereas tenure for churners (in green) is much more prevalent in the early months, particularly under 10 months and continues to decline throughout the customer lifecycle.

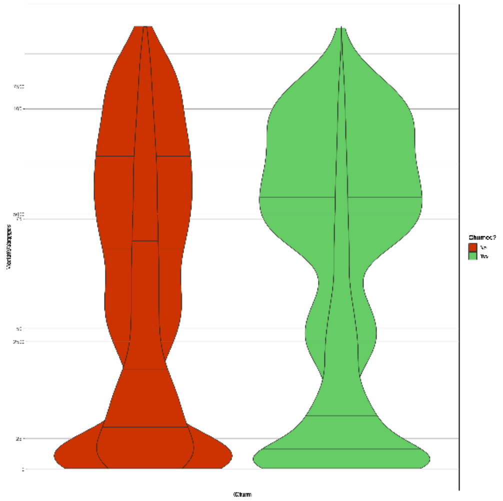


Figure 2 - Violin diagram of MonthlyCharges. Churned customers are shown in green.

Figures 2 and 3 show distinct differences in the means of both monthly charges and total charges for churners verus non-churners. T-tests on these means will help us validate if these are statistically significantly differences.

IV. T-Tests on Tenure, Monthly and Total Charges

The natural question that arises from studying the violin plots of tenure, monthly charges and total charges is whether the differences between churners and non-churner is statistically significant. Independent samples t-tests can help us determine if these visual differences are statistically significant.

I ran three independent samples t-tests, one each comparing the means of customers who churn versus those who do not churn across the integer variables tenure, monthly charges and total charges. Our hypotheses for these tests are that the true means are equal while the alternate hypotheses are that the true means are not equal.

$$H_O: Means\ for\ tenure\ are\ equal$$
$$H_a: Means\ for\ tenure\ are\ not\ equal$$

Similar hypothesis for monthly charges and total charges test for equal means with these variables. Our p-value value for tenure is $< 2.2e - 16$ rejecting our null hypothesis that the means are equal. The mean tenure for churners is ~18 months while the mean tenure for non-churners is ~37 months. This is statistically significant and what we might expect to see from this data set.

Interestingly, churners tended to have slightly higher monthly charges of ~$74/month compared to non-churners monthly charges of ~$61/month. Our t-tests revealed this difference to be statistically significant too. Churners and non-churners also have different total charges, with churners mean total charges of ~$1532 versus non-churners total charges of ~$2555. These differences also statistically significant.

V. Data Imputation Methods

Missing data – the dataset was found to have missing data in the TotalCharges variable in 11 rows/records. Other than these missing data in 11 records, all other rows were complete. While the number of records with missing data was low compared to the total number of records, three different approaches to imputing the data were performed: PMM, CART and LASSO. LASSO imputed negative values for TotalCharges, while PMM and CART imputed different values for TotalCharges. As shown in Table 1 below, PPM and CART methods for imputing missing data provided reasonable values. However, the LASSO method imputed some negative values for TotalCharges. While negative TotalCharges may be a possibility in rare circumstances, such as when overages in refunds, no other TotalCharges data exhibited negative total chares so this method was removed from further consideration.

original	PPM	CART	LASSO
NA	309.25	25.70	122.44
NA	18.90	70.15	-1,609.59
NA	216.75	19.65	-534.61
NA	63.75	67.10	-2,341.43
NA	270.95	60.00	388.41
NA	18.80	181.50	-885.68
NA	69.25	44.75	-557.52
NA	18.90	45.25	-796.42
NA	19.00	141.50	-2,128.79
NA	711.95	87.30	382.58
NA	426.65	155.65	1,078.06

Table 2 - Comparison of PPM, CART & LASSO Imputation Results

I compared the histogram of the original values of Total Charges to histograms that included the PMM and CART imputed values; there was no visible change in the histograms (likely because we were only imputing 11 values out of 7032), so PMM imputed values were chosen. Thus, I conclude the imputed values did not significantly change the distribution of Total Charges. (Note: I am not aware of a statistical technique for choosing among different sets of imputed values.)

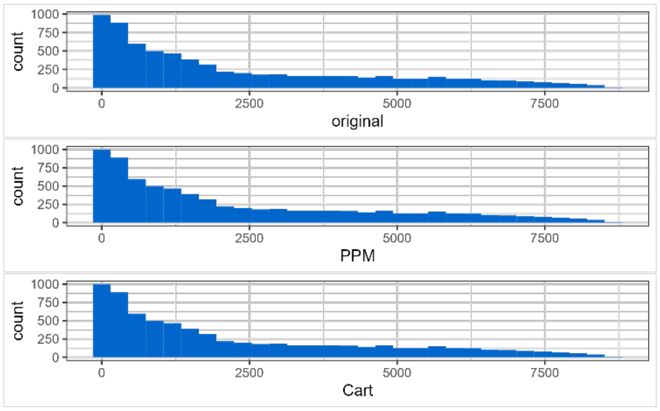


Figure 4 - Comparison of original values, PPM and CART imputed missing values for TotalCharges.

Balanced classes – In the research on classification methods, it is recommended that the classes be "balanced" – balanced meaning an equal number of cases/records/rows in each class used to classify the data. I analyzed the telco churn data set and found the balance of churn (the class variable) to be a ratio of nearly 3:1 "No" to "Yes." While much of the research literature on balanced classes referenced data sets with rare occurrences (e.g. bank fraud, cancer detection, intrusion detection) , it appears most models will perform best with balance classes. Two common methods of balancing an imbalanced dataset are under-sampling and over-sampling. For this assignment, I chose under-sampling to balance the data set to have an equal number of row/records in each class (Churn = No/Yes). This was relatively easy to implement on the telco churn data set and resulted in each class having 1,869 data points/rows/records. The models for this assignment were run on the balanced dataset.

Correlation analysis – Performing correlation analysis on categorical data is much different from the continuous data we have been using.

With categorical, especially nominal, rather than ordinal or ratio nominal data, the appropriate correlation analysis uses a measure called Cramer's V. For this data set Cramer's V was calculated and a list of those independent variables with correlation greater than 0.5 was generated for possible removal from the data set. Figure 5 show the correlation plot for all variables in the dataset.

Correlation analysis found the following strongly correlated variables:

- Total Charges is very highly correlated with several variables including partner, dependents, payment method, churn, tenure, contract and senior citizen.
- Monthly Charges is highly correlated with internet service, online backup, streaming movies, streaming tv and device protection variables.

With these data analyses completed, the data set was randomly partitioned into training and testing subsets.

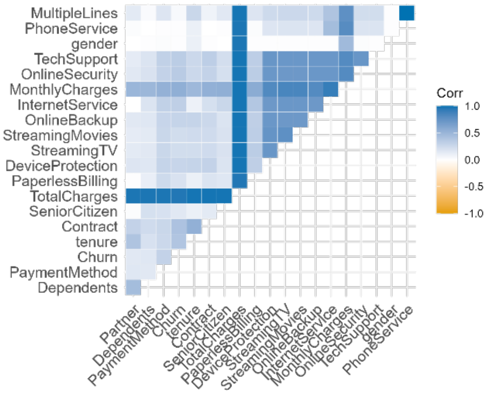


Figure 5 - Correlation Between Telco Dataset Variables

VI. Variable Selection Methods

Variable selection for data sets having nearly all their independent/predictor variables be nominal, categorical data is more complex than with continuous data. For the homework assignment, I evaluated three different approaches: a logit model with p-value based variable selection (and AIC on the overall model), chi-square variable selection using step-AIC as the criteria, and a GLM with forward- and backward-AIC as the criteria. Each approach is briefly described below.

Variable Selection Approach	Description	Results
Correlation	Remove variables that are highly correlated (based on Cramer's V) and can lead to poor predictions.	Partner, dependents, payment method, tenure, contract and senior citizen should be removed due to very high correlation with total charges. Internet service, online backup, streaming movies, streaming tv and device protection should be removed due to high correlation with monthly charges.
Logit modeling with p-value	Remove variables with p-values suggesting they are insignificant after running a logit model on the data. Logit creates dummy variables.	Logit modeling, identifying variables with (Pr>|z|)>0.5, indicates senior citizen yes, tenure, multiple lines yes, internet service provider fiberoptic, internet service no, streaming tv yes, streaming movies yes, contract one year, contract two year, paperless billing yes, payment method echeck, and total charges are the most important attributes.
Chi-square feature selection	The chi-square statistical test can be used to determine if nominal independent variables are associated with the dependent variable. Chi-square creates an attr_importance metric.	Chi-square indicates contract, online security, tech support, tenure, internet service, online backup, total charges, device protection and paperless billing are the most important attributes.

Table 3 - Summary of variable selection approaches and selected variables.

Logit model with p-value variable selection – this approach appears to be the simplest of the variable selection models, and based on the selected variables and overall model AIC, it also appears to be the lease effective approach. This method automatically creates dummy variables, and uses R's generalized linear model package to create a binomial linear model with associated p-values for each variable (dummy variables included). This approach suggests the variables SeniorCitizenYes, tenure, MultipleLinesYes, ContractOneYear, ContractTwoYear, PaperlessBillingYes, PaymentMethodElectronicCheckYes and TotalCharges are the statistically significant variables. The overall AIC of the model is ~3684. This AIC was improved by other variable selection models. Thus, this would not be a preferred approach to variable selection.

chi-square variable selection using step-AIC – Chi-square variable selection is another approach often mentioned in the research literature on nominal variable selection methods. This approach uses cross validation and linear step-AIC removing variables until the min AIC is reached. This approach suggests the variables OnlineSecurity, TotalCharges, MonthlyCharges, StreamingTV, SeniorCitizen, MultipleLines, PaymenMethod, tenure, StreamingMovies, InternetService and Contract are the most important variables with a model AIC of ~ -4712.

Chi-square, forward & backward, step-AIC – this approach uses R's stepAIC with glm() to perform forward and backward variable selection based on chi-squared. This method, like the logit approach, also creates dummy variables to represent the various options for nominal/categorical variables. The final model from this method selects SeniorCitizenYes, tenure, MultipleLinesNo, phone service, MultipleLinesYes, InternetServiceFiber optic, InternetServiceNo, StreamingTVNo internet service, StreamingTVYes, StreamingMoviesNo internet service, StreamingMoviesYes, ContractOne year, ContractTwo year, PaymentMethodCredit card, PaymentMethodElectronic check, PaymentMethodMailed check, MonthlyCharges, and TotalCharges.

While the three different approaches seem to select many similar variables, the chi-square variable selection using step-AIC is the most parsimonious and provides a robust AIC. Considering the variable selection models were run on the full set of variables, additional analysis using the same approaches could be done after removing variables shown to be highly correlated based on Cramer’s V.

VII. Modeling Methods

Several different modeling methods were used on the balanced churn data using all independent/predictor variables. These include:

1. Random forest
2. Boosting
3. Decision Tree
4. Logistic regression (stepwise)
5. Naive Bayes
6. KNN

Accuracy as a performance metric for model evaluation. In this analysis, I have decided to use accuracy as the primary measure for comparing the selected models. The reasons for choosing accuracy within the context of the telco churn analysis include:

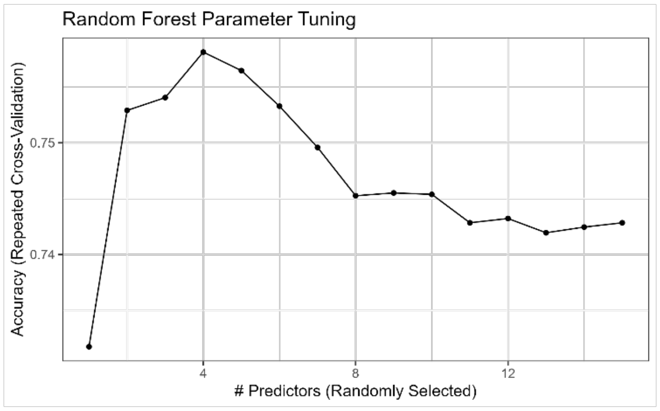
- Ease of interpretation – accuracy is an intuitive and straightforward metric which is easy to understand. It measures the proportion of correct predictions compared to the total number of predictions and provides a clear indicator of each model’s overall performance.
- Alignment with business objectives for this analysis – In analyzing churn, our goal is to maximize the overall number of correct predictions for both classes, churners and non-churners. The accuracy metric aligns closely with these objectives as it considers both True Positives and True Negatives in its calculation and presents them in an intuitive, simple metric.

In choosing accuracy as the primary performance metric, it is important to note that accuracy can be misleading if the data set is unbalanced. While the telco churn data set was initially imbalanced, all our modeling was performed on a balanced data set ensuring robust accuracy measurements. For more detail on methods tested and selected to balance the data set classes, please see the section titled *Balanced Classes*.

Confusion matrices were computed for every model, and the full detail of confusion matrix comparisons is in the *Analysis and Results* section.

Each method is briefly described below along with its testing error for comparison.

Random forest – a random forest model was developed using the full set of variables. Ten-fold cross-validation was repeated 3x (due to long computation time). Tuning of the randomForest included testing the number of variables sampled as candidates at each split (e.g. mtry) from 1 to 15 and the number of trees to grow (ntree = 500). The optimal number of predictors was found to be 4, with accuracy of 76.81%.



Gradient Boosting – A boosting model was developed using the full set of variables in the telco churn data set. Ten-fold cross validation was used during parameter tuning. Four tuning parameters were considered include the n.trees, interaction depth, minimum observations at each terminal node and shrinkage. A grid search was performed holding shrinkage low at 0.1, providing a slow learning rate to optimize the models ability to generalize, and the minimum observations per terminal node at 20, while the number of trees was evaluated from 1 to 50 and the maximum depth of each tree in the range of 1 to 10. The optimal and final values used were n.tree = 18 and interaction.depth =3, while both shrinkage and minimal observations at each terminal node were held constant at 0.1 and 20 respectively. Accuracy was 76.57% with these values.

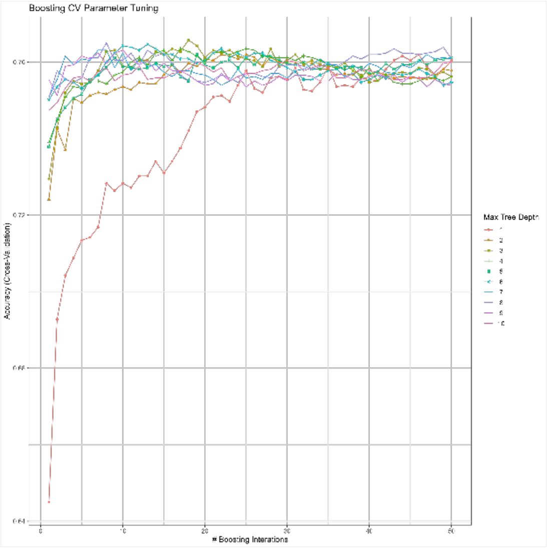
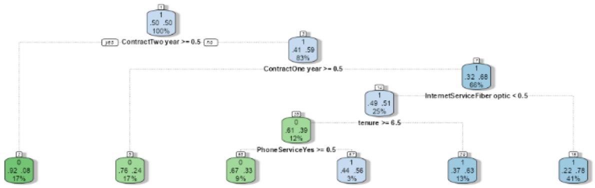


Figure 6 - Tuning of gradient boosting hyperparameters n.trees and interaction.depth.

Decision Tree – A decision tree model was develop using the full telco churn data set. The model was optimized for accuracy using repeated 10-fold cross-validation. A grid search was used to find the optimal complexity parameter in the range 0.0 to 1.0. The optimal model had a cp = 0.005 with accuracy of 75.21%. The final model is shown below.



Rattle 2023-Apr-16 19:17:56 rcw8mac

Figure 7 - Decision tree model

Logistic regression – A logistic regression model was develop using 10-fold repeated cross-validation on the reduced data set to minimize errors due to multicollinearity. The model's accuracy was 74.50%.

NiaveBayes – A Naive Bayes model was developed 5-fold repeated cross-validation on the full data set. Parameter tuning was done using a grid search across the three tunable hyperparameters of usekernal of true and false, laplace of 0 to 2.0 in increments of .2 to evaluate various levels of smoothing, and an adjust multiplier on the kernel bandwidth in the range of 0.5 to 2.0 in increments of 0.25. The model parameters with the highest accuracy were usekernel = FALSE, laplapce = 0.0 and an adjust of 0.5, resulting in accuracy of 72.69%.

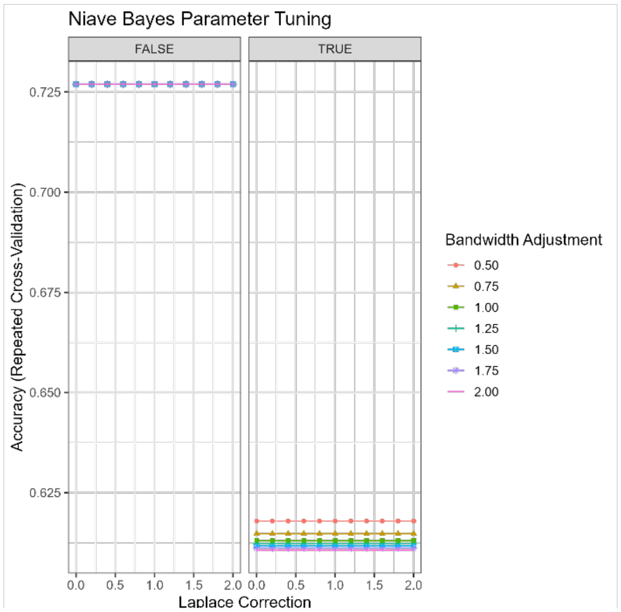


Figure 8 - Tuning of naive bayes usekernel, laplace and adjust hyperparameters.

KNN – A k-nearest neighbors' model was developed to predict churn. The data set was scaled and center prior to training the model (because the model relies on a distance measure). Ten-fold repeated cross-validation was utilized along with a grid search for hyperparameter tuning with k ranging from 1 to 50. For the final model k=23 performed optimally with an accuracy of 74.66%. Figure 8 provides a plot of the accuracy performance of various k-values.

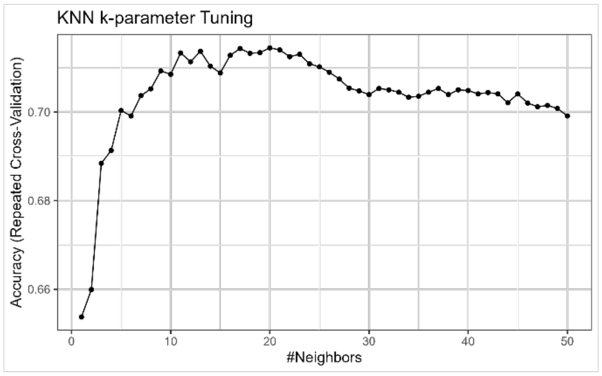


Figure 9 - K-parameter Tuning for KNN showing optimal accuracy at k = 20

VIII. Analysis and Results

In the analysis and results section, I share some insights on model performance, variable selection, and future model enhancements.

Business Impact

Across all models, the accuracy ranged from 72.69% for naïve bayes to 76.81% for the random forest model. While a 4% difference across all models seems relatively minor, considering the large difference in tenure for churners (18 months) versus non-churners (37 months) times the number of total customer relationships at a large telco times the additional total charges ($1,532 for churners vs $2,555 for non-churners), 4% additional lift could provide significant revenue.

For telecommunications business, the analysis demonstrates that reducing churn and increasing the lifetime value of their customers is achievable. Generally, it is less costly to keep existing customers than it is to acquire new customers, thus significant additional revenue is available by extending current customer tenure and these models have shown their effectiveness identifying customers who may churn. One approach to implementing these models would include running the model each month on the entire customer base/list and providing an incentive for the customers continued loyalty – perhaps in the form of discounts, special savings on desirable products, or other incentives.

For this project, the primary goal was to find a model with the highest accuracy. However, other metrics may be relevant based on business decisions. For example, if a telco decided to offer an expensive incentive in an attempt to retain customers, the precision metric may be a secondary performance metric to ensure resources are not wasted on false positives. Recall may be another performance metric in terms of capturing most of the customers who will churn, and by this measurement the random forest model performs better than most other models, except logistic regression. Table 4 summarizes various performance metrics across all of the models tested.

Performance Metric	Random Forest	Boosting	Decision Tree	Logistic Regression	Naive Bayes	KNN
Accuracy	76.81%	76.57%	75.21%	74.50%.	72.69%.	74.66%
Sensitivity	0.7300178	0.7069272	0.6767318	0.7602131	0.5825933	0.6625222
Specificity	0.8064516	0.8082437	0.8243728	0.7437276	0.8745520	0.8225806
Pos Pred Value	0.7919075	0.7881188	0.7954071	0.7495622	0.8241206	0.7902542
Neg Pred Value	0.7475083	0.7321429	0.7165109	0.7545455	0.6749654	0.7072419
Precision	0.7919075	0.7881188	0.7954071	0.7495622	0.8241206	0.7902542
Recall	0.7300178	0.7069272	0.6767318	0.7602131	0.5825933	0.6625222
F1	0.7597043	0.7453184	0.7312860	0.7548501	0.6826223	0.7207729
Prevalence	0.5022302	0.5022302	0.5022302	0.5022302	0.5022302	0.5022302
Detection Rate	0.3666369	0.3550401	0.3398751	0.3818020	0.2925959	0.3327386
Detection Prevalence	0.4629795	0.4504906	0.4272971	0.5093666	0.3550401	0.4210526
Balanced Accuracy	0.7682347	0.7575855	0.7505523	0.7519704	0.7285726	0.7425514

Table 4 - Summary of performance measures across all models.

Variable insights

The numerical/integer variables including TotalCharges, MonthlyCharges, and tenure appeared in many of the models; this might provide some insight into a very concise and parsimonious model.

Due to the existence of so many categorical variables, I found this dataset to be challenging – many of the approaches taught in OMSA focus on continuous data, and these techniques don't work easily for nominal, categorical data or require significant modification.

Model enhancements

It would interesting to investigate how MonthlyCharges or TotalCharges might be incorporated into the model to maximize the predicted revenue rather than model the predicted churn. From a business perspective, maximizing customer revenue might be better than predicting churn – although a longer customer relationship should also increase revenue.

IX. Lessons Learned

For me, this has been a fantastic class. Unlike many of the other OMSA classes that focus solely on the mathematical or technical aspects of a specific area of analytics, say time-series or regression, ISYE 7406 provided a much broader perspective of the analytical process – a more holistic view of selecting a model, tuning the models hyperparameters, measuring the performance of the model versus other models and drawing conclusions. I now understand the process of analytical modelling.

In this course, I learned about several new types of models including principal components, partial least squares, LDA, QDA, Naïve Bayes, and all the tree-based models. Ensemble models were also new to me, and very interesting.

I have a much deeper understanding of smoothing, splines and support vector machines and how they work mathematically. The course presentations are excellent and have the right mix of theory, mathematics and R examples.

There are a number of other topics that were not taught in the course, but the course format and assignments encouraged me to research them externally; these include variable selection techniques, methods of data imputation, and a better understanding of model performance metrics. I found that the homework assignments encouraged and allowed for exploration of these areas, both in terms of workload and the grading format. The course assignments are well thought out.

In addition to the course presentations and assignments (both excellent), I appreciate that Dr. Mei and the TAs made themselves available each week via Zoom. While attendance was sometimes small, these sessions contributed positively to the learning experience considerably. Finally, I hope Dr. Mei will consider a consider developing a second course, leveraging the same format, focused on large language models and generative AI.

X. References

[1] Gohel D, Skintzos P (2023). _flextable: Functions for Tabular Reporting_. R package version 0.9.0, <https://CRAN.R-project.org/package=flextable>.

[2] - Grolemund G, Wickham H (2011). “Dates and Times Made Easy with lubridate.” _Journal of Statistical Software_, *40*(3), 1-25. <https://www.jstatsoft.org/v40/i03/>.

[3] - Hutson G (2021). _OddsPlotty: Odds Plot to Visualise a Logistic Regression Model_. R package version 1.0.2, <https://CRAN.R-project.org/package=OddsPlotty>.

[4] - Kassambara A (2022). _ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'_. R package version 0.1.4, <https://CRAN.R-project.org/package=ggcorrplot>.

[5] - Kassambara A (2023). _ggpubr: 'ggplot2' Based Publication Ready Plots_. R package version 0.6.0, <https://CRAN.R-project.org/package=ggpubr>.

[6] - Kuhn M (2022). _caret: Classification and Regression Training_. R package version 6.0-93, <https://CRAN.R-project.org/package=caret>.

[7] - Majka M (2019). _naivebayes: High Performance Implementation of the Naive Bayes Algorithm in R_. R package version 0.9.7, <https://CRAN.R-project.org/package=naivebayes>.

[8] - Meyer D, Zeileis A, Hornik K (2023). _vcd: Visualizing Categorical Data_. R package version 1.4-11, <https://CRAN.R-project.org/package=vcd>. Meyer D, Zeileis A, Hornik K (2006). “The Strucplot Framework: Visualizing Multi-Way Contingency Tables with vcd.” _Journal of Statistical Software_, *17*(3), 1-48. doi:10.18637/jss.v017.i03 <https://doi.org/10.18637/jss.v017.i03>. Zeileis A, Meyer D, Hornik K (2007). “Residual-based Shadings for Visualizing (Conditional) Independence.” _Journal of Computational and Graphical Statistics_, *16*(3), 507-525. doi:10.1198/106186007X237856 <https://doi.org/10.1198/106186007X237856>.

[9] - Milborrow S (2022). _rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'_. R package version 3.1.1, <https://CRAN.R-project.org/package=rpart.plot>.

[10] - Müller K, Wickham H (2023). _pillar: Coloured Formatting for Columns_. R package version 1.9.0, <https://CRAN.R-project.org/package=pillar>.

[11] - Müller K, Wickham H (2023). _tibble: Simple Data Frames_. R package version 3.2.0, <https://CRAN.R-project.org/package=tibble>.

[12] - port SobSDiR, revised ebMM, Dutky mbS (2021). _bitops: Bitwise Operations_. R package version 1.0-7, <https://CRAN.R-project.org/package=bitops>.

[13] - R Core Team (2022). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

[14] - Romanski P, Kotthoff L, Schratz P (2021). _FSelector: Selecting Attributes_. R package version 0.33, <https://CRAN.R-project.org/package=FSelector>.

[15] - Sarkar D (2008). _Lattice: Multivariate Data Visualization with R_. Springer, New York. ISBN 978-0-387-75968-5, <http://lmdvr.r-forge.r-project.org>.

[16] - Stekhoven DJ (2022). _missForest: Nonparametric Missing Value Imputation using Random Forest_. R package version 1.5. Stekhoven DJ, Buehlmann P (2012). “MissForest - non-parametric missing value imputation for mixed-type data.” _Bioinformatics_, *28*(1), 112-118.

[17] - Therneau T, Atkinson B (2022). _rpart: Recursive Partitioning and Regression Trees_. R package version 4.1.19, <https://CRAN.R-project.org/package=rpart>.

[18] - van Buuren S, Groothuis-Oudshoorn K (2011). “mice: Multivariate Imputation by Chained Equations in R.” _Journal of Statistical Software_, *45*(3), 1-67. doi:10.18637/jss.v045.i03 <https://doi.org/10.18637/jss.v045.i03>.

[19] - Venables WN, Ripley BD (2002). _Modern Applied Statistics with S_, Fourth edition. Springer, New York. ISBN 0-387-95457-0, <https://www.stats.ox.ac.uk/pub/MASS4/>.

[20] - Waring E, Quinn M, McNamara A, Arino de la Rubia E, Zhu H, Ellis S (2022). _skimr: Compact and Flexible Summaries of Data_. R package version 2.1.5, <https://CRAN.R-project.org/package=skimr>.

[21] - Wickham H (2016). _ggplot2: Elegant Graphics for Data Analysis_. Springer-Verlag New York. ISBN 978-3-319-24277-4, <https://ggplot2.tidyverse.org>.

[22] - Wickham H (2022). _stringr: Simple, Consistent Wrappers for Common String Operations_. R package version 1.5.0, <https://CRAN.R-project.org/package=stringr>.

[23] - Wickham H (2023). _forcats: Tools for Working with Categorical Variables (Factors)_. R package version 1.0.0, <https://CRAN.R-project.org/package=forcats>.

[24] - Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.” _Journal of Open Source Software_, *4*(43), 1686. doi:10.21105/joss.01686 <https://doi.org/10.21105/joss.01686>.

[25] - Wickham H, François R, Henry L, Müller K, Vaughan D (2023). _dplyr: A Grammar of Data Manipulation_. R package version 1.1.1, <https://CRAN.R-project.org/package=dplyr>.

[26] - Wickham H, Henry L (2023). _purrr: Functional Programming Tools_. R package version 1.0.1, <https://CRAN.R-project.org/package=purrr>.

[27] - Wickham H, Hester J, Bryan J (2023). _readr: Read Rectangular Text Data_. R package version 2.1.4, <https://CRAN.R-project.org/package=readr>.

[28] - Wickham H, Vaughan D, Girlich M (2023). _tidyr: Tidy Messy Data_. R package version 1.3.0, <https://CRAN.R-project.org/package=tidyr>.

[29] - Williams GJ (2011). _Data Mining with Rattle and R: The art of excavating data for knowledge discovery_, series Use R! Springer. <https://rd.springer.com/book/10.1007/978-1-4419-9890-3>