Using Supervised and Unsupervised Machine Learning to Predict and Prevent Telecommunications Churn

Preston Cusick, Ethan Eckmann, Rian Kahlon, Daniel Wendland

School of Information, University of Texas at Austin

I 310D: Introduction to Human-Centered Data Science

Professor Abhijit Mishra

1 May 2025

The Problem and Objective

Our group has chosen to base our Human-Centered Data Science project on the Telco dataset originally provided by IBM at kaggle.com. The Telco dataset is a simulated dataset originally published by IBM, designed to reflect realistic customer churn patterns. As a telecommunications provider, the Telco company's number one priority is preventing "churn", or the rate at which customers discontinue paid services. Using Machine Learning techniques and additional analysis, our team will identify the main factors driving customer churn in the Telco company and offer solutions that would help lower the rate of customer churn in the future.

Methodology Outline

There are three main parts of our approach. By taking multiple approaches, we can present a well-rounded conclusion that will be used to guide Telco leadership on preventing churn.

1. Supervised machine learning
    a. Train a predictive Machine learning model on the dataset and test F1 scores for different Machine Learning algorithms. Choose the best algorithm and use it to create a churn "probability" score for each entry.
    b. Focusing on interpretability, identify which data fields are most influential for determining churn probability according to the model.
2. Unsupervised machine learning

Based on the churn probability associated with each entry, categorize entries that would have been churned last quarter at Telco into tiers of risk for further review and analysis.

3. Additional analysis and visualizations

Provide scatter plots and pie charts to analyse fields in the dataset for further analysis.

The Data

Data Overview

The dataset was sourced from kaggle.com, a website dedicated to sharing Datasets and Machine Learning projects. Throughout the course of our project, it was often necessary to drop or modify some fields from the original dataset due to redundancies, irrelevance, or data type errors. Considering what data to use in a Machine Learning model is always very important. Transformation mapping was also required before we could build our machine learning models due to the high amount of boolean or string type fields in the data. Whenever significant changes were made to the dataset, we created and stored CSV files for the different versions of the dataset. These can be found at our page on Data.world, link is at the end of this report.

- Telco_customer_churn.xlsx - The original dataset from https://www.kaggle.com/datasets/yeanzc/telco-customer-churn-ibm-dataset/data
- Telco_customer_churn.csv - The original dataset converted into CSV
- Telco_customer_churn_A.csv - Original dataset but with the addition of correcting a type-error with 'Total Charges' field. Was string when should be numeric
- Telco_customer_churn_B.csv - Same as Telco_customer_churn_A, but with some fields removed due to redundancies or irrelevance for the project. Fields removed: 'City', 'Zip Code', 'Churn Reason', 'CustomerID', 'Count', 'Country', 'State', 'Lat Long', 'Latitude', 'Longitude', 'Phone Service', 'Internet Service', 'Contract', 'Churn Label'
- churn_predict_A.csv - Same as Telco_customer_churn_A, but with the addition of the 'Probability' field
- churn_predict_B.csv - Same as Telco_customer_churn_B, but with the addition of the 'Probability' field

- churn_clusters_A.csv - Same as churn_predict_A, but with the addition of 'Cluster_Label' field
- churn_clusters_B.csv - Same as churn_predict_B, but with the addition of 'Cluster_Label' field

Data Fields

Dropped fields: Fields that were dropped and not used in the predictor Machine Learning model.

Used fields: Fields that were used in the predictor Machine Learning model.

Additionally, we also created two new fields: 'Probability' and 'Churn_Label'

Three data types are present in the data, String, Numerical, and Boolean

String - Words or sentences

Numerical - Data with floating points or integers

Boolean - Words or numbers, but only with two possible options. Gender was an exception and labeled as String because realistically the dataset could incorporate more than two possible gender options such as "Prefer not to say"

Dropped fields	**Used fields**
CustomerID - String A unique ID that identifies each customer *Reason for drop: Irrelevant* IDs not needed for Machine learning	Gender - String The customer's gender: Male, Female
Count - Numerical A value used in reporting/dashboarding to sum up the number of customers in a filtered set *Reason for drop: Redundant* All entries had same value of '1'	Senior Citizen - Boolean Indicates if the customer is 65 or older: Yes, No
Country - String The country of the customer's primary residence	Partner - Boolean Indicate if the customer has a partner: Yes, No

Reason for drop: Redundant All entries had same value of 'United States'	
State - String The state of the customer's primary residence *Reason for drop: Redundant* All entries had same value of 'California'	Dependents - Boolean Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc.
City - String The city of the customer's primary residence *Reason for drop: Irrelevant* We do not want to train our model to rely on customer location for predicting churn	Tenure Months - Numerical Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above
Zip code - Numerical The zip code of the customer's primary residence *Reason for drop: Irrelevant* We do not want to train our model to rely on customer location for predicting churn	Multiple Lines - Boolean Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No
Lat Long - Numerical The combined latitude and longitude of the customer's primary residence *Reason for drop: Irrelevant* We do not want to train our model to rely on customer location for predicting churn	Online Security - Boolean Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No
Latitude - Numerical The latitude of the customer's primary residence. *Reason for drop: Irrelevant* We do not want to train our model to rely on customer location for predicting churn	Online Backup - Boolean Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No
Longitude - Numerical The longitude of the customer's primary residence. *Reason for drop: Irrelevant* We do not want to train our model to rely on customer location for predicting churn	Device Protection - Boolean Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No
Phone Service - Boolean Indicates if the customer subscribes to home phone service with the company: Yes, No *Reason for drop: Redundant* Because the dataset is on existing or previous	Tech Support - Boolean Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No

Telco customers, they will all have either phone service or some type of internet service	
Internet Service - Boolean Indicates if the customer subscribes to Internet service with the company: No, DSL, Fiber Optic, Cable *Reason for drop: Redundant* Because the dataset is on existing or previous Telco customers, they will all have either phone service or some type of internet service	Streaming TV - Boolean Indicates if the customer uses their Internet service to stream television programing from a third party provider: Yes, No. The company does not charge an additional fee for this service
Contract - String Indicates the customer's current contract type: Month-to-Month, One Year, Two Year. *Reason for drop: Irrelevant* Contract type was not related to the services provided by Telco, monetary statistics, or any demographics	Streaming Movies - Boolean Indicates if the customer uses their Internet service to stream movies from a third party provider: Yes, No. The company does not charge an additional fee for this service
Churn Label - Boolean Yes = the customer left the company this quarter. No = the customer remained with the company. Directly related to Churn Value *Reason for drop: Redundant* Churn Value field serves the same purpose but in numerical format	Paperless Billing - Boolean Indicates if the customer has chosen paperless billing: Yes, No
Churn Score - Numerical A value from 0-100 that is calculated using the predictive tool IBM SPSS Modeler. The model incorporates multiple factors known to cause churn. The higher the score, the more likely the customer will churn *Reason for drop: Irrelevant* The main goal of our project is to create our own prediction model to help Telco prevent churn. Relying on data from Telco's existing prediction model would be counterintuitive	Payment Method - String Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check
Churn Reason - String A customer's specific reason for leaving the company. Directly related to Churn Category *Reason for drop: Irrelevant* Too many unique string values to be useful in the ML model	Monthly Charges - Numerical Indicates the customer's current total monthly charge for all their services from the company

	Total Charges - Numerical Indicates the customer's total charges, calculated to the end of the quarter specified above
	Churn Value - Boolean 1 = the customer left the company this quarter. 0 = the customer remained with the company. Directly related to Churn Label
	CLTV - Numerical Customer Lifetime Value. A predicted CLTV is calculated using corporate formulas and existing data. The higher the value, the more valuable the customer. High value customers should be monitored for churn.

New Data Fields

Probability - Numerical Value generated by our Supervised Machine Learning model. Ranges from 0-1. Higher values mean higher predicted probability for churn	Churn_Label - String Generated by our Unsupervised Machine Learning model based on the Probability field. Labels all entries into 4 possible tiers of risk. High, Moderate, Low, and Very Low risk

Methodology and Insights

Methodology - Supervised Prediction model

For supervised learning we tested three different models and picked the best overall performing one. We tested the four main metrics, but mainly focused on F1 score, as well as the confusion matrix. After examining the models, we will move on to visualization and explanation

First, we tested logistic regression which draws a linear line between churn and no-churn, predicting based on the weighted sum of features.

Figure 1 - Logistic Regression metrics

Metric	Score

Accuracy	79.30%
Precision	67.40%
Recall	52.30%
F1-score	58.90%

	Predicted No Churn	Predicted Churn
Actual No Churn	908	101
Actual Churn	191	209

Next, we tested the Random Forest model which uses many decision trees at the same time, combining their results through voting to improve accuracy. This model saw better overall scores then the previous but the worst recall.

Figure 2 - Random Forest metrics

Metric	Score
Accuracy	79.50%
Precision	69.20%
Recall	50.00%
F1-score	58.10%

	Predicted No Churn	Predicted Churn
Actual No Churn	920	89

Actual Churn	200	200

Finally, we tested the XGBoost model. This model builds many small decision trees sequentially, each learning from the previous tree's mistake, making it more accurate for complex patterns. This accuracy was reflected in the metrics with the best scores overall (with precision only being second to Random Forest).

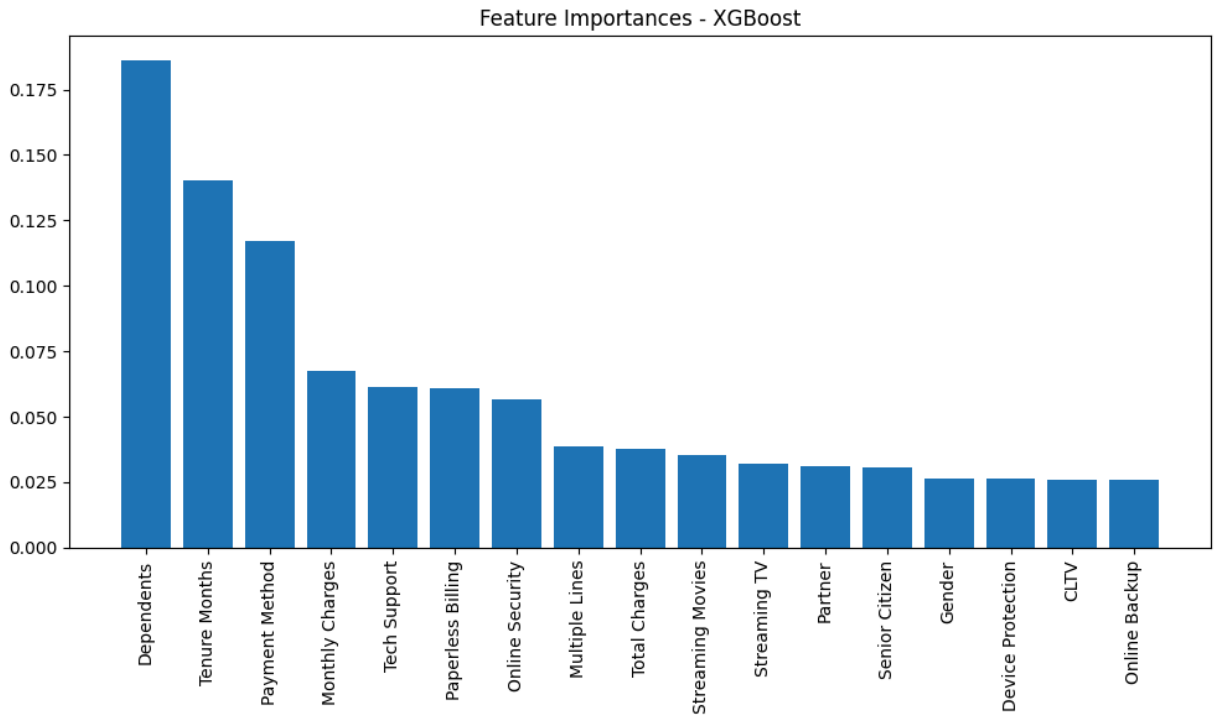
Figure 3 - XGBoost metrics (Chosen model)

Metric	Score
Accuracy	79.90%
Precision	68.60%
Recall	54.00%
F1-score	60.40%

	Predicted No Churn	Predicted Churn
Actual No Churn	910	99
Actual Churn	184	216

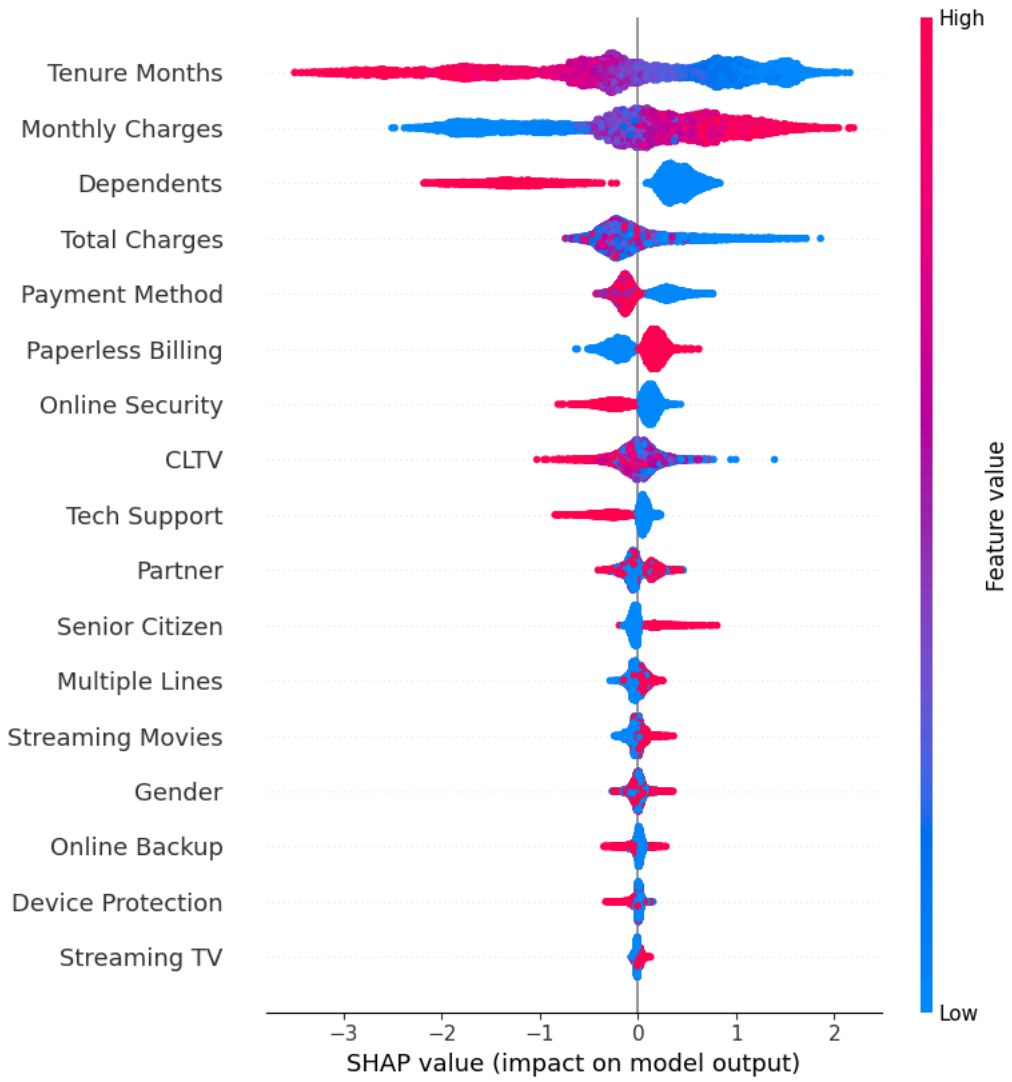
We then moved on to examine feature importance. In XGBoost, a feature importance is first calculated for an individual tree based on the value it provided. It then averages the frequency of use across all trees, regardless of performance. We found that the top 3 features were Dependents, Tenure Months, and Payment Methods.

Figure 4 - Feature Importance



We now knew what features were the most impactful, but not how they were impactful. To remedy this we utilized the SHAP library. SHAP (SHapley Additive exPlanations) is a method based on the Shapley value from game theory to explain and visualize ML models. SHAP allows us to visualize feature importance in terms of which direction they push prediction.

Figure 5 - SHAP General Explainer



Furthermore, the model above does not tell the full story of the Payment Method feature, as it is categorical in nature. So we further examine this by again using SHAP to show the impact of each value.

Figure 6 - SHAP Payment Method Explainer

```
Electronic check: Avg SHAP impact on churn = 0.3262

Mailed check: Avg SHAP impact on churn = -0.1733

Auto-pay: Avg SHAP impact on churn = -0.1243
```

Insights - Supervised Prediction model

Expanding upon the chosen model, XGBoost, shown in figure 3, we can see its strengths compared to the other two models. It had the greatest accuracy, with 79.9%, showing it predicted churn correctly most of the time. The precision, 68.6%, tells us that it was right about two thirds of the time. It also had by far the best recall of the models, at 54%, although this number is still somewhat lacking, it did correctly identify a majority of the churners. It had the best F1 score as well. Finally, looking at the confusion matrix, we can see that the model is much better at predicting those who stayed with Telco.

Then looking at figure 5, we get a better idea of how the top 3 features impact churn. To explain how to interpret this figure, the main two attributes are directions and color. Left indicates a negative push, or push towards no churn, and red shows a positive push towards churn. Red means a high feature value and blue means a low feature value. For tenure, towards the right and blue, that means that low tenure months push people towards churn. Having a dependent, red and to the left, means a lower churn risk. However, for payment methods it is more complicated than what can solely be found in this figure.

To explore the Payment Methods field's impact, we must use figure 6. There we see that both Mailed Check and Auto-pay have negative pushes from churn (i.e, customers paying with Mailed Check and Auto-pay were at lesser risk of churn), with -0.1733 and -0.1243 respectively. However, Electronic Check has a huge positive push towards churn with 0.3262. This means Electronic check payments are a major factor in contributing to customer churn.

Methodology - Unsupervised K-means clustering model

To complement our supervised churn prediction model, we also applied an unsupervised learning approach using k-means clustering. The purpose was to group customers into risk - based segments based on the churn probability scores generated from XGBoost model, the top performing supervised method, which yielded an F1- score of 60.4% By clustering these probability values, we aimed to identify meaningful risk tiers among customers who actually churned in the most recent quarter. This risk segmentation would enable telco to be able to proactively target intervention strategies for customers most likely to churn in the future.

Before applying k-means, we cleaned the data by removing entries with missing values in the probability field to ensure accuracy in clustering. We chose four clusters to create distinct and manageable tiers: High risk, Moderate risk, Low risk, and Very Low risk. These tiers reflect the probability distribution and help translate abstract model outputs into actionable business categories. We exclusively did the clustering on customers who churned, in order to retroactively evaluate whether our model could have identified at risk individuals.

Insights - Unsupervised K-means clustering model

Of the customers who were churned in the last quarter, the unsupervised k-means model grouped them into the following results:

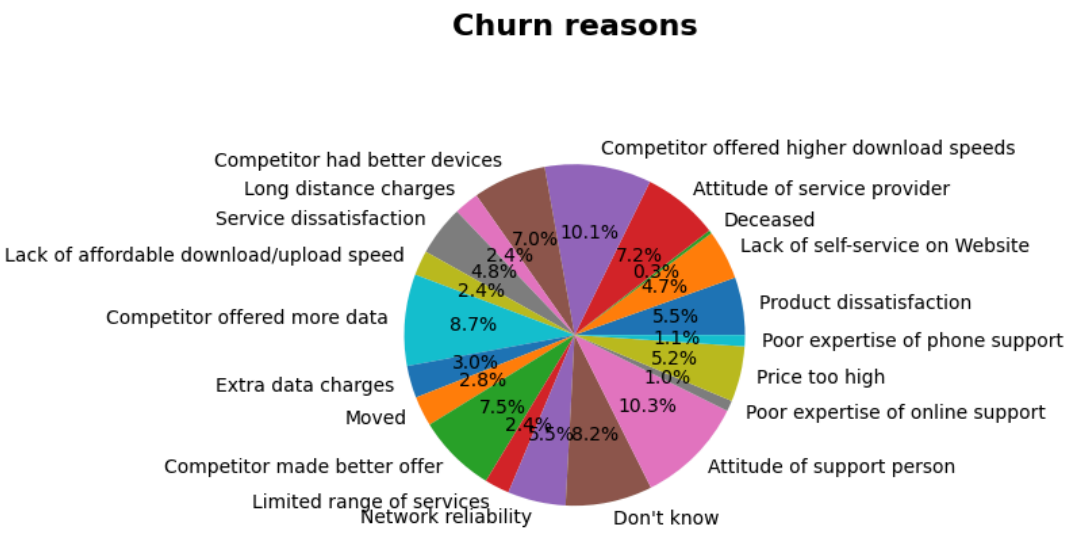
- High Risk: 595 customers (31.83%)
- Moderate Risk: 576 customers (30.82%)
- Low Risk: 438 customers (23.43%)
- Very Low Risk: 260 customers (13.91%)

The average predicted churn probability for all customers churned in the last quarter was about 0.596. Of the customers that were churned in last quarter, the average customer would have been determined as being slightly more at risk of churn than not. These results support the idea that the K-means clustering model would be useful for creating effective targeted intervention plans.

Additional analysis and visualizations

In addition to creating our machine learning models, we also did analysis on the content of the dataset itself and the relationships between the fields in the dataset and our generated churn 'probability' field.

Figure 7 - Churn Reasons Pie chart

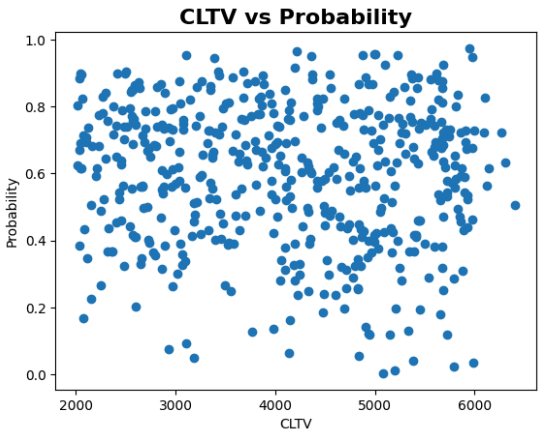


Displaying the different reasons for customer churn allows us to see some additional weak points for Telco's customer retention. The top three reasons for churn were:

- 'Attitude of support person'
- 'Competitor offered higher download speeds'
- 'Competitor offered more data'

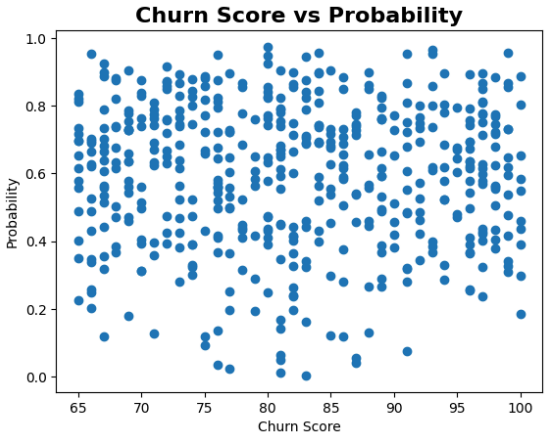
In addition to the findings of our machine learning models, this basic visualization helps us to find other problems Telco must face to bring down churn rates.

Figure 8 - CLTV vs Probability of churn Scatter plot



The results of correlation tests between Customer Life-Time Value and Probability features demonstrated low linear relationships with churn because their Pearson's r value equated to about -0.107. This means that high Customer Life-Time Value is not necessarily the most important aspect one might think of when considering churn probability.

Figure 9 - Churn Score vs Probability of churn Scatter plot



The results of correlation tests between Churn Score (Value generated by Telco's own risk prediction model) and Probability features demonstrated extremely low linear relationships with churn because their Pearson's r value equated to about -0.021. From this we can conclude that Telco's method of generating churn risk prediction for their customers is vastly different from our model. Considering that our model has a F1-score of 60.4%, this would suggest that Telco's risk predictions are highly inaccurate.

Conclusion

We found the features that were most related with churn prediction were tenure months, whether a customer has dependent(s), and customer payment method. High tenure months, having a dependent, and paying with either Mailed check or Auto-pay were strong indicators against customer churn. However, using electronic checks was a strong indicator towards churn. We propose that Telco focus on retaining customers with dependents and high tenure months

through programs like loyalty rewards. Telco should also encourage new or existing customers to opt for Auto-pay methods.

By including the K-means clustering model in our project, we transitioned from simple churn prediction to risk-level segmentation that provides a better framework for deciding which customers to focus on. The high-risk cluster contained 31.83% of customers churned last quarter. Meanwhile, the moderate-risk cluster contained 30.82%. If Telco had been utilizing our models, these groups would have been flagged and focused efforts would have been made by the company to reduce the churn rate. In the future, Telco should use this model to easily find their most at-risk customers and take steps to prevent these individuals from churning. The model demonstrates real-world application and shows how predictive analytics can enhance current real-time customer retention programs.

To recap, we recommend that Telco leadership should:

- Focus on retaining existing customers and targeting new customers that have dependents
- Increase loyalty programs
- Encourage Auto-Payment
- Reorganize the customer support department
- Offer higher download speeds and data than competitors
- Utilize both demonstrated machine learning models together to predict the risk levels of their customers and categorize them into manageable tiers of risk

Appendix

In reflection of this project we faced three major limitations, size of the dataset, amount of factors, and time constraints. While the data set was large, it was still somewhat small for a proper ML model. This mixed with the imbalance found in the data set and us only using around 20% of the original in actual testing left room for improvement. Next was the original scope of our project versus the number of factors found. To explain, while we accomplished the original goal of finding what features had the biggest impact on churn, the accuracy and detail behind this was still left with many unanswered variables. These variables include potential bias, unexplored features, and overall ML fairness. This leads to our final constraint of time which ties back to what we wished to discover in our original scope and contrasts it with all the questions we were left with.

Next we will address the questions we received on our presentation.

- How is feature importance computed in XGBoost?

  This was already discussed earlier during the methodology section in regards to figure 4. To summarize, XGBoost calculates importance based on how often a feature is used in trees and how much it improves model gain.

- Could the model show bias? What factors cause it?

  The model could show bias towards non-churners. This is due to non-churners significantly outnumbering churners in the data. It is possible that due to this imbalance the model chose to favor non-churners as they were the majority value.

In the future we would try to address this using a variety of techniques such as SMOTE (Synthetic Minority Oversampling Technique) to ensure that minority classes are not overlooked.

- How does random state affect reproducibility?

  We made sure to set the same random state for each of the tested models to ensure reproducibility for those running our code. This was important due to us using multiple models and so many features,

- What steps would you take for future improvement?

  In reference to our limitation and previous questions, in the future we would have greater focus on feature engineering and model fairness. In terms of feature engineering we casted quite a wide net with the number of features used in our models. We did this as we wanted to look into as many potential churn causes as possible, but model accuracy may have suffered as a result. Doing more to improve metrics such as recall would be a priority in the future. As for model fairness, as mentioned previously the dataset used was imbalanced. Ideally in the future we would take further precautions to avoid potential model bias either through different training techniques or an improved dataset. Lastly, we would take a greater look into model fairness to ensure that minority groups are not being targeted. We chose not to closely examine such features due to time

constraints, but such features should be examined before any further actions by Teclo are taken.

References

mikesuperman. (2023, March 9). *SHapley Additive exPlanations or SHAP: What is it ?* Data Science Courses | DataScientest. https://datascientest.com/en/shap-what-is-it

Brownlee, J. (2020, January 16). *SMOTE for Imbalanced Classification with Python*. Machine Learning Mastery. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

Abid Ali Awan. (2023, June 28). *An Introduction to SHAP Values and Machine Learning Interpretability*. Datacamp.com; DataCamp. https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability

Brownlee, J. (2016, August 30). *Feature Importance and Feature Selection With XGBoost in Python*. Machine Learning Mastery. https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/

Github link

https://github.com/E-Eckmann/I310D-Telco-project-EERKPCDW

Data.world link

https://data.world/bethew/i310d-telco-project-eerkpcdw-datasets