Theran Meadows
Project 3: Churn Prediction
Milestone 3
August 10, 2024

Topic

My project will be to use a data set provided by IBM for a fictional company,Telco, to determine who of their customers are more likely to churn. From the same data set, I would also like to investigate what characteristics or reasons of the long term customers prevented them from leaving.

Business Problem

When a customer churns from a company, it means that that customer is no longer a customer of that company's services. When customers churn, it can be for many different reasons. Maybe the product stopped being reliable. Maybe the product got too expensive and was no longer worth it in the eyes of the customer or, even worse, competition from another company has attracted them to try their similar product.

In my own life, I know I have churned a couple different times for my car insurance, who I had streaming platform subscriptions with, or even where I ate my fast food at. No worries, there will always be a new customer to take a churned customer's place right? Yes and no. Yes because there are probably so many different people and types of audiences that a product could go to to attract new customers. No, because it is actually more expensive to sell to a new customer than it is to sell to an existing customer. If you take on a lot of customers but also have a high churn rate, you won't have enough money to sustain a constant flow of customers coming in and going out.

There needs to be a balance of trying to achieve low churn and maintain existing customer satisfaction. My business questions to answer will be:

What customers are more likely to churn and why?

Datasets

Kaggle.com is providing the data set I will be using. It is a dataset given by IBM for a fictional company called Telco. It has over 7000 rows of customer data including tenure, the products they have or don't have, how much they spend, and some personal information about them. This is a fictional set, any personal identifiable information is a farce.

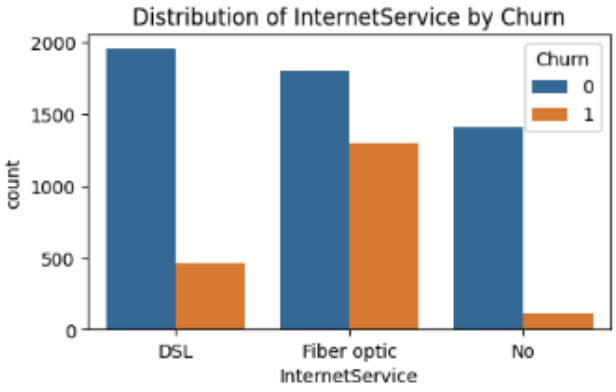
Methods

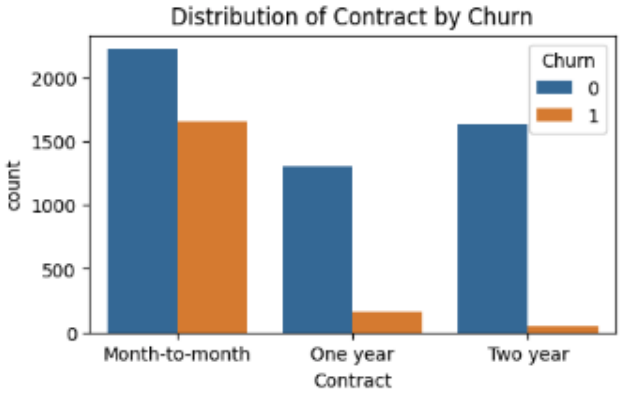
I will be doing exploratory data analysis as well as 2 predictive models. I would like to try a logistic regression model and a Naive Bayes model. Logistic regression model is easy to interpret and can classify who churned and who didn't.

A Naive Bayes model would also be good because it does well with multiple classifications and can classify customers as 'low', 'medium', or 'high' risk.

Analysis

Exploratory Data Analysis



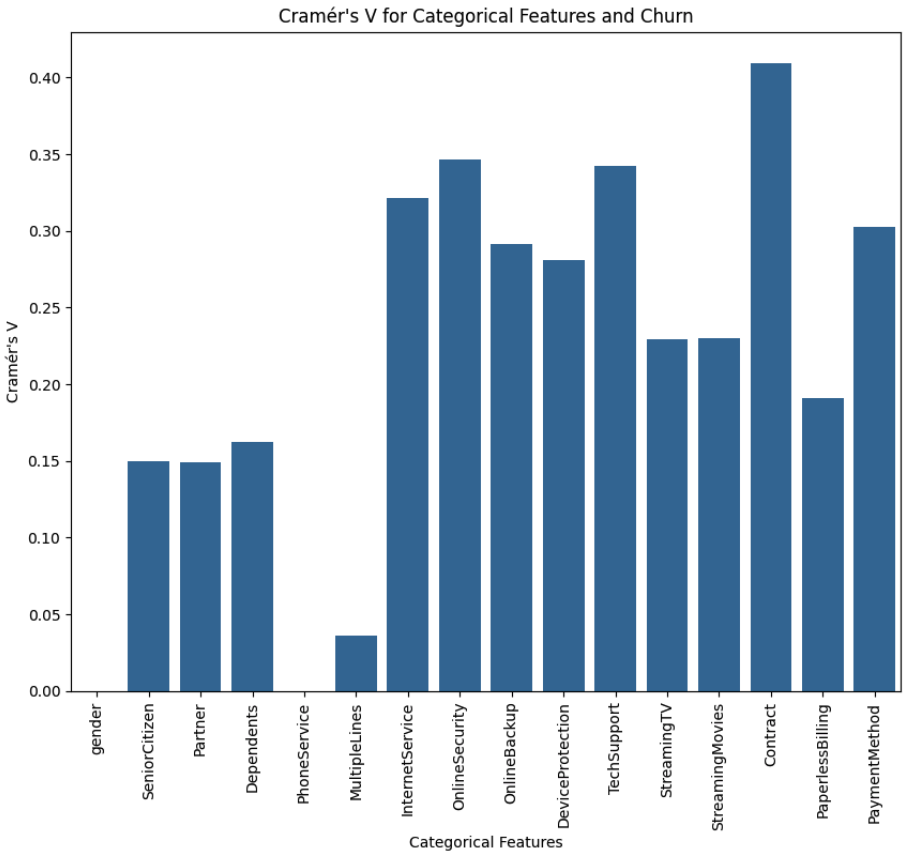


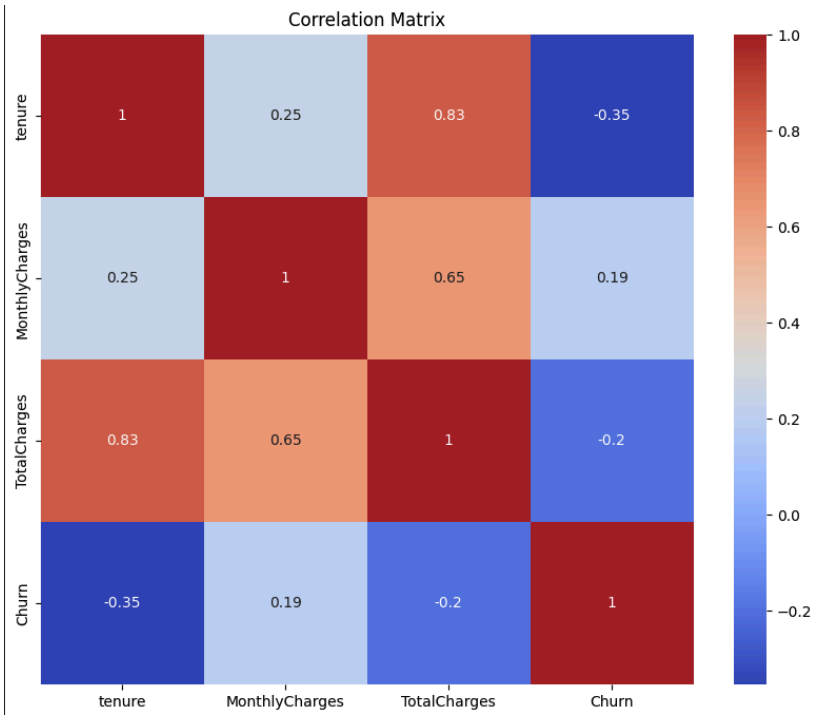
Of the categorical features:

['gender', 'SeniorCitizen', 'Partner', 'Dependents', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod']

Having a fiber optic cable or a contract seems to be what customers had before they churned the next month.

I also did a Cramers V statistic to give a better overview visual of the categorical features and how they correlate with churning. Online security, device protection, and payment method also seem to have some significance. Though In my opinion, I really couldn't see why these would have significance as security and device protection seem to be products and payment methods seem to be a personal preference.





Conclusion

In conclusion, preventing customer churn requires a multifaceted approach that includes improving customer experience, personalizing engagement, offering flexible options, and proactively addressing potential churn indicators. By leveraging these insights, businesses can develop targeted strategies to retain at-risk customers and continue to onboard new customers. This analysis is just an example of many that could help a potential company in need. Predictive models like Random Forest and Naive Bayes provide valuable tools for understanding and mitigating churn.

Assumptions/Limitations

Churning prediction is specific to each company. Although principles and trends can be seen across multiple companies, there are many different reasons both financial and non financial reasons as to why a customer might cancel services. I am limited in just this dataset for this specific company.

It would also be hard to know if the reason there is a high churn rate is because of something political and not something to do with the services of the company. For example, if the company came out to the public about being a supporter of LGBTQ+ communities, customers that don't support those communities might leave simply for having a different political standpoint.

Ethical Considerations

No ethical considerations to consider since the data I am using is open to the public and is fictional.

Challenges/Issues

This is my first time analyzing any data set to investigate churning rates. This data set is specific to this company only. Although there may be companies with similar products, this data set should only be considered for this project.

Future Uses

Churn prediction is different for every company. It should be up to the company to do a periodic pull of all customers and then analyze those that left. If the churn is significant, then it might be wise to allocate some money into churn prevention and customer retention.

Recommendations

From this data set, offering incentives or discounts to customers with month-to-month contracts or high monthly charges could reduce churn. Additionally, improving customer service for those using electronic check payments and enhancing features for long-term contract customers could further decrease churn rates.

References

BlastChar. (2018, February 23). *Telco customer churn*. Kaggle.

https://www.kaggle.com/datasets/blastchar/telco-customer-churn

10 Questions

1. **What are the most significant factors that contribute to customer churn?**

The most significant factors contributing to customer churn in the Telco dataset include contract type, monthly charges, tenure, and payment method. Customers with month-to-month contracts, higher monthly charges, shorter tenure, and electronic check payment methods are more likely to churn.

2. **How did you handle missing data in the dataset?**

Luckily, there were no missing values in this data set. Had there been, depending on the field, I may have used the median value for the column.

3. **Why did you choose the specific models for this analysis?**

I chose Random Forest for its ability to handle complex interactions and non-linear relationships in the data. Additionally, Naive Bayes was selected for its effectiveness with categorical data and simplicity. I wanted two to have validation between the two.

4. **How do you evaluate the performance of your predictive models?**

The performance of the models was evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. These metrics provide a comprehensive understanding of the model's performance, especially in handling imbalanced datasets like churn prediction.

5. **How do the categorical features correlate with churn?**

Categorical features such as contract type, payment method, and internet service type show a strong correlation with churn. For instance, customers on month-to-month contracts and those using electronic check payment methods have higher churn rates. Cramér's V was used to measure the strength of association between these categorical features and churn.

6. **What preprocessing steps were taken before training the models?**

Preprocessing steps included converting categorical variables to numerical values using Label Encoding and standardizing numerical features to ensure they were on the same scale.

7. **How does the churn rate vary across different customer segments?**

The churn rate varies significantly across different customer segments. For example, customers with month-to-month contracts have a higher churn rate compared to those with one-year or two-year contracts. Similarly, customers with shorter tenure (less than 12 months) and higher monthly charges are more likely to churn. Maybe they got a shorter contract because they weren't sure if they would want to stay.

8. **Did you use any techniques to handle imbalanced data?**

Yes, techniques such as oversampling (using SMOTE) and undersampling were used to handle the imbalanced data. Additionally, performance metrics like precision, recall, and ROC-AUC were used to evaluate the model, focusing on its ability to correctly identify churners despite the imbalance.

9. **How does the model handle new data?**

This is a fictional data set but had it been real, a pipeline to regularly update the model with new data and re-evaluate its performance metrics to ensure it adapts to any changes in customer behavior would help significantly as the company grows. Stakeholders will be better prepared to make decisions that could prevent churning early.

10. **What are the business implications of your findings?**

The findings have significant business implications. By identifying the key factors that drive churn, the company can develop targeted retention strategies. For example, offering incentives or discounts to customers with month-to-month contracts or high monthly charges could reduce churn. Additionally, improving customer service for those using electronic check payments and enhancing features for long-term contract customers could further decrease churn rates.