Sarah Theriot

Project 2

Milestone 2

Keeping Customers Subscribed: Predicting Netflix Churn Data

Topic

Keeping Customers on Board: A Data-Driven Look at Netflix Churn

This project looks at how Netflix can use customer data to understand why people cancel their subscriptions. The goal is to find patterns in behavior that could help predict when someone might leave. With this kind of insight, streaming platforms could step in earlier and offer solutions that keep people interested. Whether it's a discount, a new feature, or better content recommendation, small changes might make a big difference in keeping customers from walking away.

Business Problem

Streaming platforms like Netflix depend on monthly subscriptions to make money, so when users cancel, it really hurts. Even though some cancellations are expected, too many can be a sign that something’s wrong. It might mean people aren’t finding shows they like, or maybe they think the price isn’t worth it anymore. As someone studying data science, I wanted to explore how companies like Netflix can use data to figure out what’s pushing customers away. If we can understand the warning signs, we might be able to predict who’s about to leave and why. That gives companies a chance to respond before it’s too late. The goal of this project is to build a model that can help spot those patterns and give teams the info they need to keep users happy and subscribed.

Dataset

The dataset I used for this project comes from Kaggle and includes information about 5,000 different Netflix users. Each row represents one customer and shows things like their gender, age, subscription plan, how long they've been using the service, and whether they've churned or stayed. It also includes data on things like how often they contact customer service and what devices they use to stream. It's not real Netflix data, but it's built to be very similar, and it works well for a project like this. Even though it's synthetic, it still gives a good idea of what kinds of patterns could exist in real life. One reason I liked this dataset is that it's already pretty clean, which made it easier to jump right into the analysis. It also has a mix of both numerical and categorical features, which helps when trying to build a prediction model.

Methods

I started by exploring the dataset using summary statistics and visualizations to get a feel for what the data was showing. I looked at how many customers had churned versus stayed, which plans were most popular, and whether age or device type seemed to play a role in who leaves. I created bar charts, pie charts, line graphs, and scatter plots to help spot patterns. After getting a better understanding of the data, I chose logistic regression as the main model for predicting churn. Logistic regression is a good starting point for classification problems like this one, especially when you want to understand what features are most important. I used scikit-learn in Python for the modeling part and split the data into training and test sets to check how well the model performs.

Analysis

The biggest thing that stood out from the visuals is that most customers in the dataset didn’t churn, but there are some clear patterns among the ones who did. People with the basic plan are much more likely to leave than those on standard or premium. That could be because they aren’t getting enough value or maybe the content feels limited. Age also seems to matter. Younger customers, especially those in their late teens and twenties, have higher churn rates. This might be because they’re more budget-conscious or tend to switch platforms more often.

Device type was another interesting one. People using phones or tablets seem more likely to churn than those using TVs or laptops. That might mean the experience on smaller devices isn’t as strong, or maybe those users are more casual and less committed. When I ran the logistic regression model, plan type, age, and monthly usage came out as some of the biggest predictors of whether someone would cancel. The model had decent accuracy and gave good insight into which customers are most at risk. That could help the company target outreach or offer deals before someone leaves.

Assumptions and Limitations

One big assumption in this project is that the synthetic dataset is close enough to real Netflix data to be useful. Even though it was built to reflect realistic patterns, I can’t know for sure how well it matches actual customer behavior. I also assumed that patterns like younger users canceling more often or people with basic plans being more likely to leave would hold true in real life. That might not always be the case.

Another limitation is that the data doesn’t include things like customer reviews, satisfaction scores, or notes about why someone canceled. Those things could offer important context. It also doesn’t capture outside factors like competition from other streaming services, price changes, or personal events like job loss. The model can only work with the data it has, so if anything important is missing, that could weaken the predictions. Still, I think this project offers a helpful place to start when it comes to understanding and preventing churn.

Challenges

One of the biggest challenges in this project was making sure the visualizations said something useful. With so many variables in the dataset, it was easy to get lost in the details or make graphs that looked nice but didn’t really help answer my main question. I had to keep reminding myself that the goal wasn’t just to explore the data, but it was to find patterns that could actually help Netflix keep more of their customers.

Another challenge was trying to avoid bias when interpreting the results. For example, just because younger users were canceling more often, that doesn’t mean age is the main reason. There could be other factors, like money, time, or even interest in certain types of content. It was hard not to jump to conclusions, especially when the graphs seemed to show strong patterns. I also found it tricky to balance being detailed with keeping things readable. It’s easy to get technical in data science, but I wanted to make sure the story stayed clear and grounded.

Recommendations

Based on what I found, Netflix and similar streaming services could focus on a few key areas to keep customers from canceling. First, paying attention to how often customers log in is important. If someone hasn’t used their account in a while, it might be a good time to send them a reminder or offer a special deal to bring them back. Second, encouraging more engagement by promoting content that matches users’ favorite genres or showing personalized recommendations could help people watch more and feel like the service is worth it. Third, looking at subscription plans, offering flexible options or discounts for people with basic plans might make a difference since those users tend to cancel more. Finally, companies should keep collecting and analyzing data regularly. The more they understand why people leave, the better they can design their offers and improve the experience for all users.

Implementation Plan

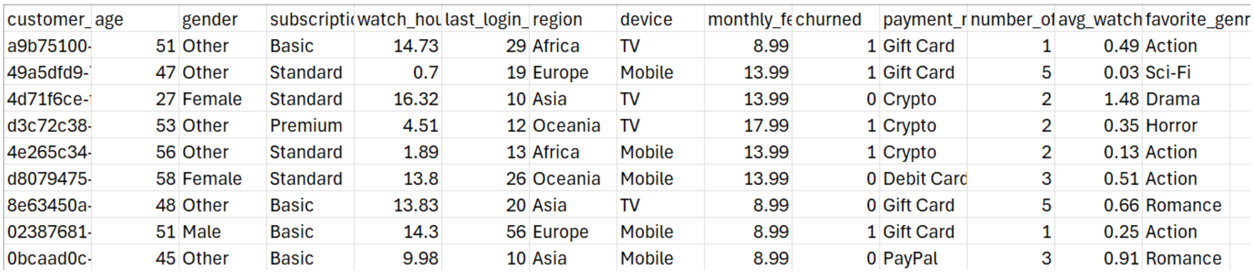
One way Netflix could use this project is by setting up a system that flags customers who look like they might cancel soon. The company could then reach out with special offers or personalized messages to remind those users why they like the service. This system would need to update regularly with new data so it stays accurate. Another idea is to add this kind of churn prediction into their existing customer service tools. That way, support staff could see who is at risk and focus their efforts where it matters most. It might also be helpful to run small tests to see which offers or messages work best for different types of customers. Overall, this plan would help Netflix use data smartly to keep customers happy and reduce cancellations before they happen.

Ethical Assessment

Working with customer data comes with a lot of responsibility. Even though this project uses fake data, in the real world, companies have to protect people's privacy and make sure they don't misuse sensitive information. It's important that customers know how their data is being used and that it's handled fairly. Predicting who might cancel could lead to better service, but it could also be used in ways that feel invasive or unfair, like targeting certain groups too much or making assumptions about people based on incomplete data. That's why any company using models like this needs to be careful and transparent about what they're doing. For me, this project is about showing how data can help improve the experience for everyone without crossing ethical lines. It's a reminder that behind every number is a real person, and that should always be kept in mind.

Appendix

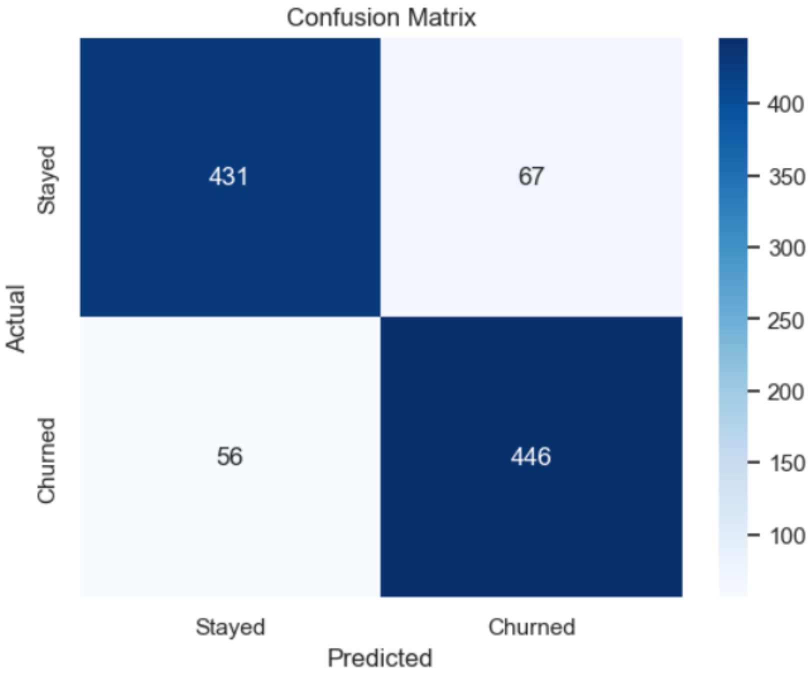
Dataset:



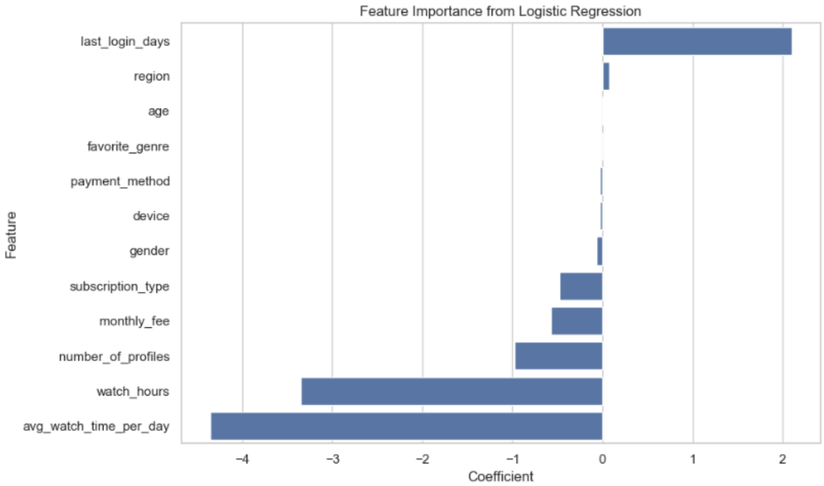
customer_	age	gender	subscripti	watch_hou	last_login_	region	device	monthly_fe	churned	payment_r	number_ol	avg_watch	favorite_genr
a9b75100-	51	Other	Basic	14.73	29	Africa	TV	8.99	1	Gift Card	1	0.49	Action
49a5dfd9-	47	Other	Standard	0.7	19	Europe	Mobile	13.99	1	Gift Card	5	0.03	Sci-Fi
4d71f6ce-	27	Female	Standard	16.32	10	Asia	TV	13.99	0	Crypto	2	1.48	Drama
d3c72c38-	53	Other	Premium	4.51	12	Oceania	TV	17.99	1	Crypto	2	0.35	Horror
4e265c34-	56	Other	Standard	1.89	13	Africa	Mobile	13.99	1	Crypto	2	0.13	Action
d8079475-	58	Female	Standard	13.8	26	Oceania	Mobile	13.99	0	Debit Carc	3	0.51	Action
8e63450a-	48	Other	Basic	13.83	20	Asia	TV	8.99	0	Gift Card	5	0.66	Romance
02387681-	51	Male	Basic	14.3	56	Europe	Mobile	8.99	1	Gift Card	1	0.25	Action
0bcaad0c-	45	Other	Basic	9.98	10	Asia	Mobile	8.99	0	PayPal	3	0.91	Romance

Visualizations:

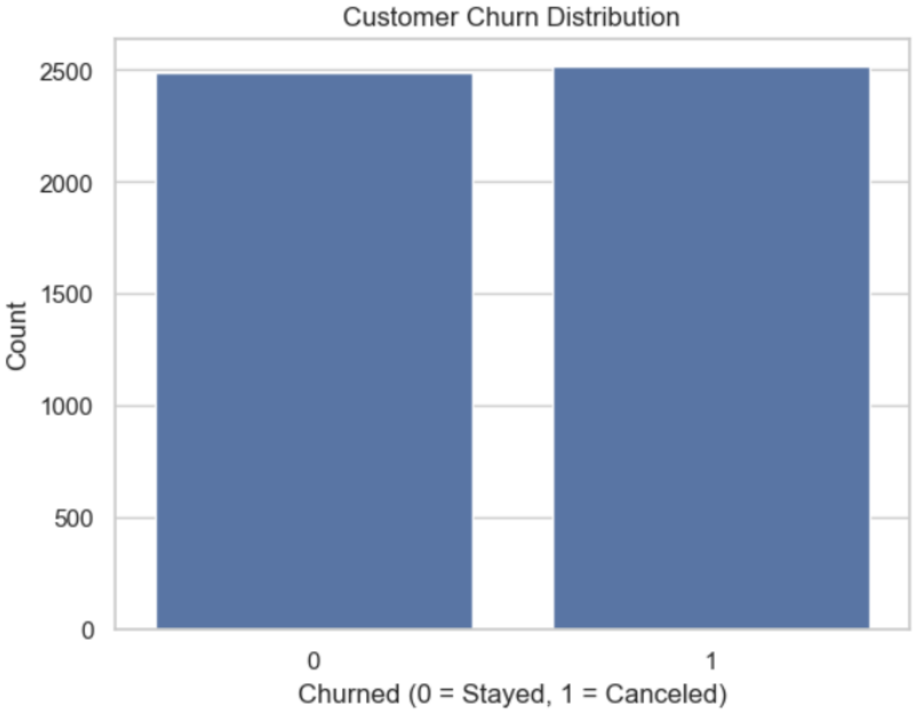
Visual A:



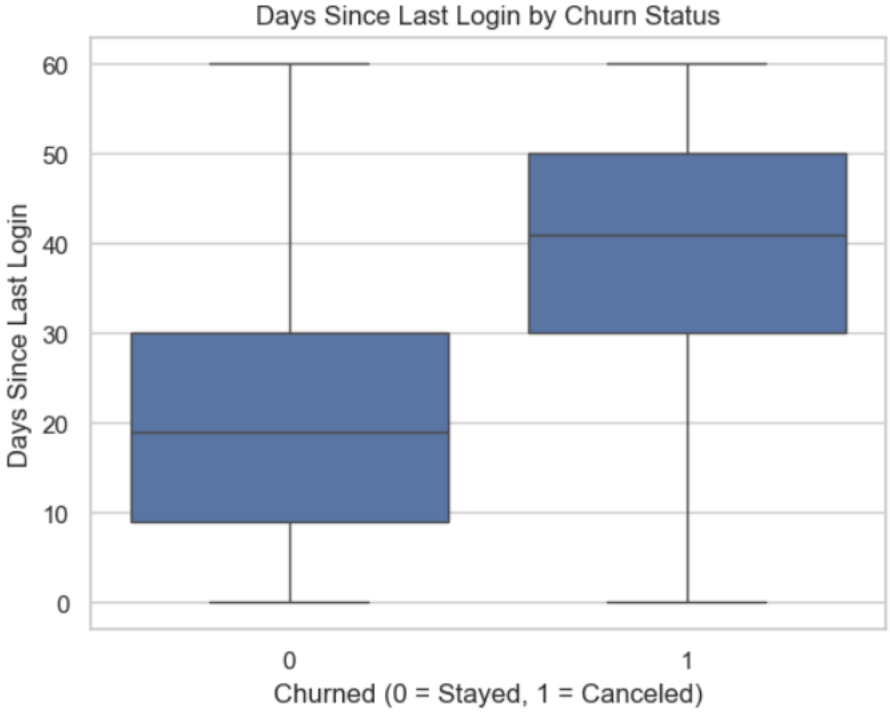
Visual B:



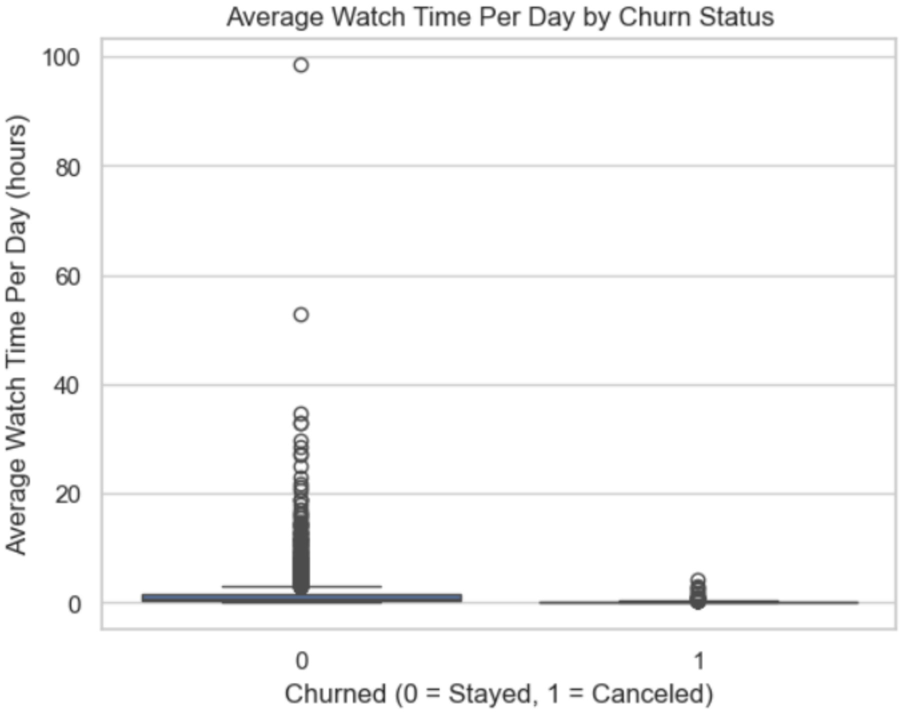
Visual C:



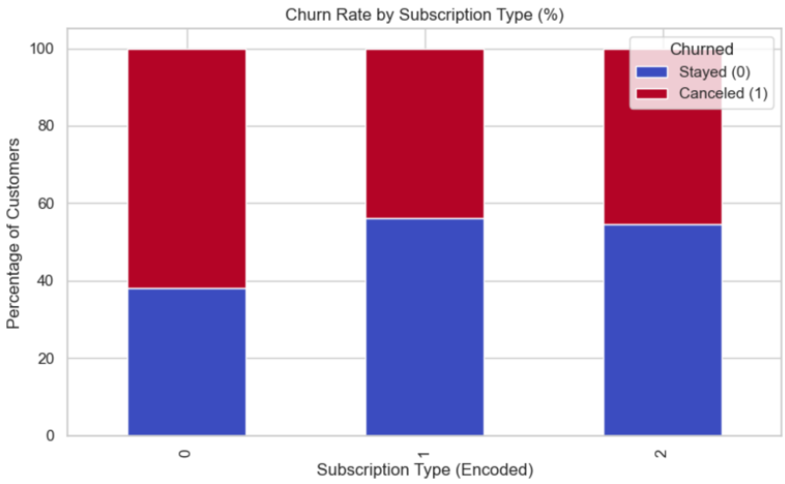
Visual D:



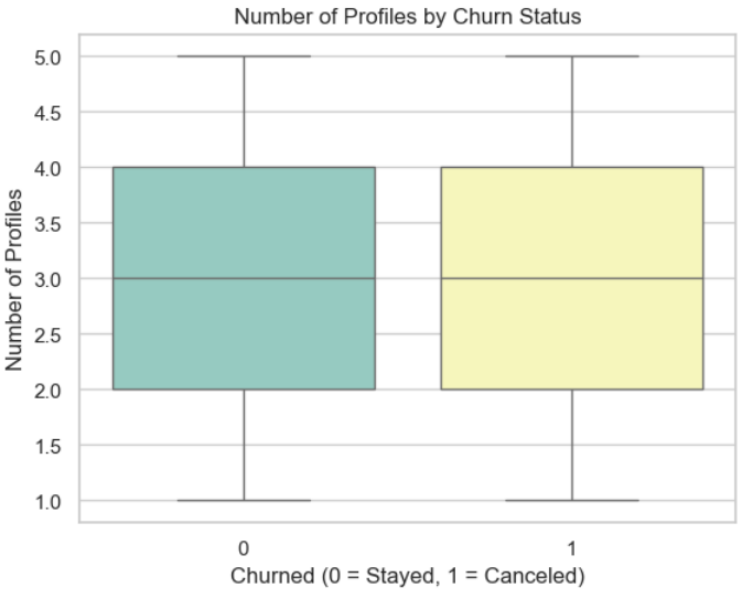
Visual E:



Visual F:



Visual G:



References

Wadood, A. (2023). Netflix Customer Churn Dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/abdulwadood11220/netflix-customer-churn-dataset

10 Audience Questions:

1. How accurate is your churn prediction model?
2. What features were most important in predicting churn?
3. Why did you choose logistic regression over other models?
4. How does the synthetic nature of the data affect your results?
5. Could the model be biased against certain groups?
6. How would Netflix use these predictions in practice?
7. Are there any risks in using customer data for churn prediction?
8. What additional data would improve your model?
9. How often should the model be updated?
10. What are the limitations of your analysis?