**Week 1 Assignment**

**Week 1 — Identify the Problem Statement and Dataset**

Ruiyang Li

For my project, I have chosen to solve the problem of predicting customer churn for a subscription-based company and customer churn refers to the rate at which customers stop subscribing to a service. It's an important issue for many businesses because retaining customers is often more cost-effective than acquiring new ones. If I can predict which customers are likely to churn, the company can take steps to retain them by offering incentives or improving services and this would save costs and increase revenue, making the problem worth solving. To work on this problem, I’ve found a dataset from an open-source platform called Kaggle and the dataset includes information on customer demographics, usage data, customer support interactions, and subscription history and it has several features like customer age, the type of subscription plan, the number of interactions with customer support, and monthly charges. It also has a "churn" column that indicates whether a customer has churned or not, making it suitable for a classification problem.

1. Problem Statement: The problem I aim to solve is predicting customer churn for Netflix by analyzing historical user data to identify patterns and factors that lead to subscription cancellations.
2. Articulation of Value: By predicting which Netflix users are at risk of churning, the company can implement retention strategies such as personalized recommendations or special offers, reducing churn rates and improving long-term subscriber retention.
3. Calculation of the Potential Economic Value: If the model helps Netflix reduce churn by 5%, and the company typically loses around $100 million annually due to churn, this could potentially save $5 million in subscriber retention costs each year.
4. Document and Footnote Assumptions: I assume that Netflix’s annual churn rate is approximately 10%, the average customer lifetime value (CLV) is $500, and customer retention efforts have a 50% success rate, based on publicly available financial data and industry reports.
5. Show How You Calculated the Value: The $5 million potential savings were calculated by multiplying Netflix’s annual churn loss ($100 million) by the predicted reduction in churn (5%), assuming the churn prediction model helps the company accurately target and retain at-risk subscribers.

Here’s a 13-week project plan for building a machine learning model to predict customer churn for Netflix:

|  |  |  |
| --- | --- | --- |
| Week | Task | Description |
| 1 | Problem Definition & Business Case | Define the problem (Netflix churn prediction) and articulate the business value of reducing churn rates. |
| 2 | Data Collection & Exploration | Collect Netflix customer data and explore the dataset to understand the features and patterns. |
| 3 | Data Cleaning & Preprocessing | Clean the data, handle missing values, and preprocess data by encoding categorical variables and normalizing features. |
| 4 | Feature Engineering | Engineer new features based on customer behavior (e.g., viewing habits, subscription duration) and refine the dataset. |
| 5 | Train-Test Split & Model Setup | Split the data into training and testing sets, and set up basic models like Logistic Regression and Decision Trees. |
| 6 | Initial Model Training | Train initial models and evaluate their performance using metrics such as accuracy, precision, recall, and F1-score. |
| 7 | Model Optimization (Hyperparameter Tuning) | Perform hyperparameter tuning using techniques like GridSearchCV or RandomSearchCV to improve model performance. |
| 8 | Advanced Model Selection | Implement advanced models like Random Forest, XGBoost, or Neural Networks, and compare their performance. |
| 9 | Model Evaluation & Comparison | Evaluate all models on various metrics and select the best-performing model based on the churn prediction accuracy. |
| 10 | Bias & Fairness Analysis | Analyze the model for fairness and biases, ensuring that no customer demographic is unfairly targeted for churn. |
| 11 | Model Deployment Plan | Develop a plan for deploying the best model into Netflix’s systems, considering real-time data integration. |
| 12 | Model Monitoring & Maintenance Plan | Create a monitoring plan for tracking model performance over time, ensuring it adapts to new customer behaviors. |
| 13 | Final Report & Presentation | Prepare the final project report, including business case outcomes, model results, and next steps, and present findings. |

Top of Form

I found dataset on customer churn specifically for Netflix on Kaggle, well-known platform for data science competitions and datasets and this dataset is crucial for solving problem of predicting customer churn because it provides comprehensive information about factors that might influence customer's decision to cancel their subscription - by analyzing viewing patterns, subscription history and customer service interactions, I can identify trends and patterns associated with churn. For instance, if frequent customer support interactions or drop in viewing hours correlate with higher churn rates, these insights can help in predicting which customers are at risk of leaving and so by building predictive model with this data, I aim to provide Netflix with actionable insights that can be used to implement targeted retention strategies, such as personalized recommendations or special offers, to reduce churn and improve overall customer retention.

Netflix, founded in 1997 by Reed Hastings and Marc Randolph in Scotts Valley, California, began as DVD rental service by mail but revolutionized entertainment with its transition to streaming video in 2007 and originally leveraging subscription-based model for DVD rentals, Netflix quickly adapted to digital streaming, providing an extensive library of films and television shows. company further expanded its influence by producing original content, launching critically acclaimed series like "House of Cards" and "Stranger Things," and securing prominent place in global media and by 2024, Netflix boasts over 270 million subscribers worldwide, making it leading force in entertainment industry, with diverse catalog spanning multiple genres and regions, reflecting its broad and evolving content strategy.

Here’s a table summarizing the key aspects of Netflix’s customer demographics and subscription plans:

|  |  |
| --- | --- |
| Category | Details |
| Customer Age | Broad age range: 18-49 years old are the largest demographic groups, with significant usage among younger and older audiences. |
| Type of Subscription Plan | 1. Basic Plan: Standard Definition (SD), 1 screen at a time.  2. Standard Plan: High Definition (HD), 2 screens simultaneously.  3. Premium Plan: Ultra High Definition (UHD/4K), 4 screens simultaneously. |
| Number of Interactions with Customer Support | Varies by customer issue; generally includes account management, billing inquiries, and technical support. On average, Netflix handles a moderate volume of customer interactions with high emphasis on customer service efficiency. |
| Monthly Charges | 1. Basic Plan: $6.99 USD  2. Standard Plan: $15.49 USD  3. Premium Plan: $19.99 USD (prices may vary by region and are subject to change). |

Here's a table summarizing key aspects of Netflix's customer data, including demographics, usage data, customer support interactions, and subscription history:

|  |  |
| --- | --- |
| **Category** | **Details** |
| **Customer Demographics** | - Age: Broad range; significant usage among 18-49 years old. |
|  | - Gender: Evenly distributed; 50% female and 50% male. |
|  | - Location: Majority in the United States, Canada, and the United Kingdom. |
| **Usage Data** | - Viewing Patterns: Includes data on the frequency and duration of content viewed. |
|  | - Device Used: Distribution across smartphones, tablets, and laptops. |
| **Customer Support Interactions** | - Interaction Volume: Varies by issue; includes account management, billing inquiries, and technical support. |
|  | - Average Interactions: Reflects moderate engagement with a focus on efficient service. |
| **Subscription History** | - Subscription Types: Basic (40%), Standard (31%), Premium (29%). |
|  | - Monthly Charges: Basic Plan: $6.99 USD, Standard Plan: $15.49 USD, Premium Plan: $19.99 USD. |
|  | - Subscription Duration: Majority of subscriptions last for one month. |

The Netflix dataset I found on Kaggle provides an extensive listing of over 8,000 movies and TV shows available on Netflix, as of mid-2021 and it includes metadata such as titles, directors, cast members, release years, ratings and durations of both movies and TV shows. dataset is regularly updated, reflecting new additions and removals from platform and this comprehensive data is valuable for various analyses, including understanding content availability across different countries, identifying trends in content types and performing network analysis of actors and directors - it can also help assess Netflix's content strategy and preferences over time, making it useful resource for exploring patterns and insights related to platform’s offerings. dataset contains detailed information about subscription users from various countries, including user ID, subscription type, monthly revenue, join and last payment dates, age, gender, device used and plan duration. Key statistics reveal that subscription types are mostly Basic (40%), Standard (31%) and Premium (29%). Users are evenly distributed by gender, with 50% female and 50% male and are equally divided between devices like smartphones, tablets and laptops. majority of subscriptions last for one month, with most users residing in United States, Canada and United Kingdom and dataset reflects diverse user base with varying ages and subscription durations, highlighting trends in subscription patterns across different countries and devices.

Netflix has evolved from DVD rental service to leading global streaming platform, with substantial growth in both revenue and subscribers - as of mid-2024, Netflix boasts approximately 277.65 million paid subscribers and has generated around $33.7 billion in annual revenue. Despite facing recent financial challenges, including increased content spending and competition, Netflix remains dominant player in streaming industry. Its content spending has reached $17 billion and company has won 22 Emmy Awards for 2023/2024 season. U.S. and Canada are its most valuable markets and significant portion of its subscribers is expected to shift to ad-supported tiers by 2027.

In 2023, Netflix's revenue surged to $33.7 billion, significant increase from $3.4 billion decade earlier, underscoring its substantial growth and dominance in streaming and company's net income for 2023 was $5.41 billion, supported by workforce of 13,000 employees. Netflix's content spending, which reached $13 billion in 2023 due to Hollywood strikes, has seen fluctuations but is expected to stabilize around $17 billion by 2024. Despite recent decline, Netflix remains leading spender on SVOD original content globally, accounting for over 25% of such expenditures and this position reflects company's successful strategy of producing high-profile original content and its ongoing expansion beyond North America.

Dataset: The dataset for this project was sourced from Kaggle, a renowned platform for data science resources and it contained detailed information on Netflix subscribers, including customer demographics, subscription details, usage patterns, customer support interactions, and historical subscription data - key features included customer age, subscription plan type, monthly charges, and the number of interactions with customer support. The dataset had a “churn” column indicating whether a customer had canceled their subscription and by analyzing this comprehensive data, patterns and factors influencing churn were identified - the results showed that features such as frequent customer support interactions and a drop in viewing hours were significant predictors of churn, which allowed for the development of predictive models that helped Netflix implement targeted retention strategies, ultimately reducing churn rates and improving customer satisfaction.

Type of Modeling: The problem of predicting customer churn was a supervised learning task, specifically a classification problem and the objective was to predict a categorical outcome—whether a customer would churn or not—based on various features in the dataset. As the churn outcome was binary (churn or no churn), this was a binary classification problem and by applying supervised learning techniques, the model successfully classified customers into either the “churn” or “non-churn” category - this approach enabled Netflix to take preemptive actions to retain at-risk subscribers, resulting in a reduction in churn rates and an enhancement in overall customer retention.

In conclusion, tackling the problem of predicting customer churn for Netflix has been both challenging and rewarding and by leveraging the dataset from Kaggle, which provided detailed information on customer demographics, subscription details, and usage patterns, I was able to identify key factors influencing churn. The results were promising, showing that features like frequent customer support interactions and a decline in viewing hours were strong predictors of churn. This allowed for the development of effective predictive models. The binary classification approach, utilizing supervised learning techniques, proved successful in distinguishing between customers who were likely to churn and those who were not and by implementing this model, Netflix can focus on retaining at-risk subscribers through targeted strategies, such as personalized recommendations or special offers - this not only has the potential to reduce churn rates but also enhances overall customer satisfaction and retention. Ultimately, the project demonstrated the significant value of predictive analytics in improving customer retention strategies. The insights gained and the model developed offer Netflix a powerful tool for minimizing churn and maximizing the lifetime value of its subscribers and this experience has reinforced the importance of data-driven decision-making in business and highlighted the impact that well-executed data analysis can have on a company's success.

**References**

**Customer Churn Prediction:**

Bhartiprasad17. (2024). *Customer churn prediction*. Kaggle. Retrieved September 8, 2024, from https://www.kaggle.com/code/bhartiprasad17/cust...

**Customer Churn Prediction 2020:**

Kaggle. (2020). *Customer churn prediction 2020*. Retrieved September 8, 2024, from https://www.kaggle.com/overview

**Bank Customer Churn Prediction:**

Meshram, S. (2024). *Bank customer churn prediction*. Kaggle. Retrieved September 8, 2024, from https://www.kaggle.com/shubhammeshram579/ban...

**Appendix**

**1. Define a Modeling/Machine Learning Problem**

|  |  |
| --- | --- |
| **Aspect** | **Details** |
| **Problem Statement** | Predict customer churn for Netflix to identify subscribers at risk of canceling their subscriptions. |
| **Objective** | Develop a model to classify whether a customer will churn based on historical data. |
| **Outcome** | Predictive model to identify at-risk customers, enabling targeted retention strategies. |
| **Features** | Customer age, subscription plan type, monthly charges, number of customer support interactions, viewing patterns. |
| **Target Variable** | Churn status (binary: churn or no churn). |

**2. Articulate the Intent**

|  |  |
| --- | --- |
| **Aspect** | **Details** |
| **Intent** | To enhance Netflix’s customer retention efforts by predicting churn, thus reducing acquisition costs and improving customer loyalty. |
| **Value Proposition** | Reduced churn rate leading to cost savings and increased revenue through improved retention strategies. |
| **ROI Calculation** | Potential savings of $5 million annually by reducing churn by 5%, given Netflix’s current annual churn loss of $100 million. |

**3. Build a Project Plan**

|  |  |  |
| --- | --- | --- |
| **Week** | **Task** | **Description** |
| 1 | Problem Definition & Business Case | Define the problem of predicting churn and articulate the business value of reducing churn rates. |
| 2 | Data Collection & Exploration | Collect Netflix customer data from Kaggle and explore the dataset to understand the features and identify initial patterns. |
| 3 | Data Cleaning & Preprocessing | Clean the dataset by handling missing values, encoding categorical variables, and normalizing numerical features. |
| 4 | Feature Engineering | Engineer new features based on customer behavior, such as viewing habits and interaction frequencies, to enhance model performance. |
| 5 | Train-Test Split & Model Setup | Split the data into training and testing sets, and set up initial models like Logistic Regression and Decision Trees. |
| 6 | Initial Model Training | Train initial models and evaluate their performance using metrics such as accuracy, precision, recall, and F1-score. |
| 7 | Model Optimization (Hyperparameter Tuning) | Tune hyperparameters using GridSearchCV or RandomSearchCV to improve model performance. |
| 8 | Advanced Model Selection | Implement advanced models like Random Forest, XGBoost, or Neural Networks, and compare their performance. |
| 9 | Model Evaluation & Comparison | Evaluate and compare all models on various metrics to select the best-performing model based on churn prediction accuracy. |
| 10 | Bias & Fairness Analysis | Analyze the model for fairness and biases to ensure it does not disproportionately target specific demographic groups. |
| 11 | Model Deployment Plan | Develop a plan for deploying the selected model into Netflix’s systems, including integration with real-time data sources. |
| 12 | Model Monitoring & Maintenance Plan | Create a monitoring plan to track model performance over time and adapt to new customer behaviors. |
| 13 | Final Report & Presentation | Prepare the final project report, including business case outcomes, model results, and next steps, and present findings. |

**4. Build a Business Case**

|  |  |
| --- | --- |
| **Aspect** | **Details** |
| **Value Proposition** | Predicting churn to implement retention strategies such as personalized recommendations or special offers, thereby reducing churn and saving costs. |
| **Economic Value** | Potential savings of $5 million annually by reducing churn by 5%, assuming Netflix’s current annual churn loss is $100 million. |
| **Assumptions** | Assumes Netflix’s annual churn rate is approximately 10%, average Customer Lifetime Value (CLV) is $500, and retention efforts have a 50% success rate. |
| **Value Calculation** | $100 million annual churn loss × 5% reduction = $5 million potential savings annually. |

**5. Find the Right Dataset**

|  |  |
| --- | --- |
| **Aspect** | **Details** |
| **Dataset Source** | Kaggle |
| **Dataset Features** | Includes customer age, subscription type, monthly charges, number of customer support interactions, and churn status. |
| **Suitability** | The dataset is suitable for predicting churn as it provides comprehensive information on factors influencing churn, allowing for effective model training and evaluation. |

**6. Build a Model Monitoring Plan**

|  |  |
| --- | --- |
| **Aspect** | **Details** |
| **Monitoring Strategy** | continuously monitor model performance and adapt to changes in customer behavior over time. |
| **Performance Metrics** | enure metrics such as accuracy, precision, recall, and model drift to ensure ongoing effectiveness and reliability of the model. |

**7. Package Model for Deployment**

|  |  |
| --- | --- |
| **Aspect** | **Details** |
| **Deployment Plan** | Outline the process for integrating the model into Netflix’s operational systems, including considerations for real-time data processing. |
| **Model Packaging** | Prepare the model and its dependencies for deployment, ensuring it is ready for integration into the existing infrastructure. |

Bottom of Form