95-851 Making Products Count: Data Science for Product Managers

HW2: Prediction and Clustering – Fall 2023

Due September 27, 2023

Overview

Knet-Flicks, a video streaming and internet company, has provided a dataset of customers with various account details and whether they churned or not. In this assignment, you will explore patterns in customer churn and build models to predict churn risk. Understanding churn drivers can help Knet-Flicks retain valuable customers.

The use of ChatGPT or similar generative AI technology is not permitted for this assignment.

Data

The data dictionary explains the measures in the dataset. Please review them before proceeding ahead. (Do not include the target Churn (viewer_status) in the features).

Measure	Description
viewer_id	A unique ID that identifies each customer
Gender	The viewer's gender: Male, Female
Age	The customer's current age, in years, at the time the fiscal quarter ended (Q2 2022)
Married	Indicates if the customer is married: Yes, No
number_of_family_dependents	Indicates the number of dependents that live with the customer (dependents could be children, parents, grandparents, etc.)
City	The city of the customer's primary residence in California
Zip Code	The zip code of the customer's primary residence
Latitude	The latitude of the customer's primary residence
Longitude	The longitude of the customer's primary residence
Number of Referrals	Indicates the number of times the customer has referred a friend or family member to this company to date
Tenure in Months	Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above
Offer	Identifies the last marketing offer that the customer accepted: None, Offer A, Offer B, Offer C, Offer D, Offer E
Internet Service	Indicates if the customer subscribes to Internet service with the company: Yes, No
Internet Type	Indicates the customer's type of internet connection: DSL, Fiber Optic, Cable (if the customer is not subscribed to internet service, this will be None)

Avg Monthly GB Download	Indicates the customer’s average download volume in gigabytes, calculated to the end of the quarter specified above (if the customer is not subscribed to internet service, this will be 0)
Online Security	Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No (if the customer is not subscribed to internet service, this will be No)
Online Backup	Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No (if the customer is not subscribed to internet service, this will be No)
Device Protection Plan	Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No (if the customer is not subscribed to internet service, this will be No)
Premium Tech Support	Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No (if the customer is not subscribed to internet service, this will be No)
Streaming TV	Indicates if the customer uses their Internet service to stream television programing from a third party provider at no additional fee: Yes, No (if the customer is not subscribed to internet service, this will be No)
Streaming Movies	Indicates if the customer uses their Internet service to stream movies from a third party provider at no additional fee: Yes, No (if the customer is not subscribed to internet service, this will be No)
Streaming Music	Indicates if the customer uses their Internet service to stream music from a third party provider at no additional fee: Yes, No (if the customer is not subscribed to internet service, this will be No)
unlimited_hi-res_streaming	Indicates if the customer has paid an additional monthly fee to have unlimited hi-res streaming: Yes, No
subscription_type	Indicates the customer’s current contract type: Month-to-Month, One Year, Two Year
Paperless Billing	Indicates if the customer has chosen paperless billing: Yes, No
Payment Method	Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check
Monthly Charge	Indicates the customer’s current total monthly charge for all their services from the company
Total Charges	Indicates the customer’s total charges, calculated to the end of the quarter specified above
Total Refunds	Indicates the customer’s total refunds, calculated to the end of the quarter specified above
Total Extra Data Charges	Indicates the customer’s total charges for extra data downloads above those specified in their plan, by the end of the quarter specified above
Total Long Distance Charges	Indicates the customer’s total charges for long distance above those specified in their plan, by the end of the quarter specified above
Total Revenue	Indicates the company's total revenue from this customer, calculated to the end of the quarter specified above (Total Charges - Total Refurnds + Total Extra Data Charges + Total Lond Distance Charges)
viewer_status	Indicates the status of the customer at the end of the quarter: Churned, or Stayed

It is OK to leave out some features, if you think they that are not that relevant. In such a case, just explain what you are doing and note your assumptions.

Tasks

You should turn in a Jupyter notebook with code and visualizations to answer:

1. Perform exploratory analysis on the data and prepare it for modeling. Examine summary statistics, handle missing values/outliers, make binary and dummy variables where applicable. Also make sure to normalize the numeric variables before feeding into the models. (5 points)

2. Explore if there are segments of customers with similar account profiles and churn rates. Use clustering and explain your choice of number of clusters. (5 points)
3. Build two models to predict churn using different supervised learning approaches. What are the top 5 most predictive factors? How do the models compare in terms of accuracy? (6 points)
4. Based on the models, which customer segment in the data is most at risk for churn? What actions would you recommend to reduce churn for this segment? (4 points)

Rubric Points Distribution:

Exploratory data analysis and data preparation	5
Development of clustering model	3
Determination of number of clusters	2
Development of first supervised learning model predicting churn	2
Development of second supervised learning model predicting churn	2
List of 5 most important factors influencing churn	1
Evaluating accuracy of the two supervised learning models	1
Identifying segment at highest risk of churn	2
Recommendations for reducing churn for the segment	2

Hints/Suggestions:

1. Describe how you have chosen to handle outliers and missing values. Think before dropping the columns completely or imputing the nulls - which makes more sense? Refer to data for product management content for EDA, handling missing data, feature selection, etc.
2. Include visualizations such as histograms and boxplots where appropriate in describing the results of EDA
3. Describe how the findings from EDA influence your choice of data preparation and choice of modeling techniques
4. Consider ways to reduce dimensions of data suitable for clustering, remember to scale your data before doing that - provide reasons for using scaling technique. Mention your reasons for using a particular clustering algo, and rationale behind choosing number of clusters. Market segmentation and unsupervised learning lecture notes are helpful.
5. Do not rename datafiles
6. For prediction, DO NOT forget to do train_test_split, provide your reasoning for choosing the appropriate metric of your choice and mention which model performs better - (overfitting/underfitting/residuals/etc.). But first think about if it is a regression or a classification problem before proceeding.
7. For the supervised learning models, consider how you will select the features and measure accuracy of the resulting models. How can you avoid overfitting?
8. Based on feature importance of model results (google up how to do this, there could be multiple methods), provide your recommendations in the end.

Submission:

The submission will be a Jupyter Notebook with the name DSPM_HW2_<your Andrew id>.ipynb, and an html export of the notebook.