Project Title

Customer Segmentation and Churn Prediction in Telecommunications Companies

CS5228 Course Mini Project, Feb 2026

Date: Winter 2026, SoC, NUS, as a part of CS5228 Course curriculum

Introduction and Goals

During this teamwork, you see the magic of data and data mining. In today's data-driven world, telecommunications companies face the challenge of retaining customers while maximizing their business potential. Customer segmentation and churn prediction are two critical tasks that leverage data mining techniques to enhance customer experience and prevent revenue loss. This is not a kind of competition towards having higher accuracy or F1-score, but to show you have understood the challenges that a data analyst faces, and how to handle them. Telephone service companies, Internet service providers, pay TV companies, insurance firms, and alarm monitoring services often use customer attrition analysis and customer attrition rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one. Companies from these sectors often have customer service branches that attempt to win back defecting clients, because recovered long-term customers can be worth much more to a company than newly recruited clients.

Companies usually make a distinction between voluntary churn and involuntary churn. Voluntary churn occurs due to a decision by the customer to switch to another company or service provider; involuntary churn occurs due to circumstances such as a customer's relocation to a long-term care facility, death, or the relocation to a distant location. In most applications, involuntary reasons for churn are excluded from the analytical models. Analysts tend to concentrate on voluntary churn because it typically occurs due to factors of the company-customer relationship, which companies control, such as how billing interactions are handled or how after-sales help is provided.

Objective: This project traces two main objectives in the fields of unsupervised and supervised learning, respectively.

- **Unsupervised Learning:** Segment customers based on their usage patterns to identify distinct groups/clusters.
- **Supervised Learning:** Predict customer churn using historical data.

  - These 2 parts convey 20 marks out of 25 total marks of the MP
- **Extra Part:** Generative AI to enrich the dataset
  - This part conveys 5 marks

Dataset:

- **Telecom Customer Churn Dataset:** This dataset includes customer demographics, account information, and usage metrics. We can presume that this dataset shows customers' usage patterns.
  - Available at: https://drive.google.com/drive/folders/1Gmav50TkZSo7lpabcE8pD15mzuYg2dOl?usp=sharing
  - Total number of data samples: 2666 + 667 = 3333 (churn-bigml-80.csv and churn-bigml-20.csv, respectively)
  - Number of columns: 20, including 19 features and a target variable.
  - An 80% - 20% random splitting was applied to generate training and testing subsets. (Again, churn-bigml-80.csv and churn-bigml-20.csv)

Project Steps:

1. **Data Preprocessing:**
   - Load the dataset using Python libraries such as Pandas.
   - Handle missing values and perform data cleaning, if necessary.
   - Encode categorical variables using techniques like one-hot or unique integer encoding.
   - Normalize numerical features if necessary. Use the training subset as the reference.
   - You may find more about the dataset in Appendix 1 at the end of this document.

2. **Exploratory Data Analysis (EDA):**
   - Use the training subset.
   - Visualize the distribution of features using Matplotlib or Seaborn.
   - Analyze correlations between features and the target variable (churn).
   - Analyze the correlations between all 19 input/independent features. Is there any co-linearity or redundancy?

- Think! How can you find the most significant 2 (and 3) features regarding the classification task and churn classes?
- Wherever we say 'analyze', somehow a kind of visualization or print should be there to report the results of that analysis.

3. **Unsupervised Learning - Customer Segmentation:**
   - Use the training subset of data.
   - Apply clustering algorithms such as K-Means, and DBSCAN to group customers based on usage patterns.
   - Ignore the churn variable in this stage.
     1. K-Means: Determine the optimal number of clusters using the Elbow Method or Silhouette Score. After finding the optimal K, do the clustering. Try to realize what that clustering suggests. E.g., what are the similarities within a cluster? i.e., Interpret and profile each cluster to understand customer segments.
     2. DBSCAN: Try to find a good set of hyperparameters. Then see what can be summed up from the results of clustering. Interpret and profile each cluster to understand customer segments.
   - Can you find anything considerable about the distribution of churn=True/False in different clusters?

4. **Supervised Learning - Churn Prediction:**
   - Select and train 3 classification models, like Logistic Regression, Decision Trees, KNN, Random Forests, or anything else to predict churn.
   - Models are optional. A trial and error or literature review can be used to select your 3 models.
   - Evaluate model performance using standard metrics, accuracy, precision, recall, and the F1-score.
   - Calculate and report the confusion matrix, too.
   - For your 3 selected models, try to optimize the hyperparameters. A short explanation of how hyperparameters were optimized would be nice.
   - In hyperparameter optimization, only the training subset can be employed.
   - Report both training and testing performance based on optimal hyperparameters. Avoid both underfitting and overfitting. ROC analysis is optional.

5. **Generative AI** to enrich the training dataset:

- Although Gen-AI is usually being used to generate unstructured data, it can generate structured datasets, too. There are several algorithms and models developed in the field of Gen-AI to generate structured documents, here your churning data.
- You are free to use any Gen-AI algorithm, library, or model that you may find appropriate for your MP.
- Only focus on your training dataset, churn-bigml-80.csv, and employ this in that synthetic data generation process whenever it is needed.
- Try to generate up to 500 synthetic data samples. Save them as '**churn-gen-ai-test-data.csv'**
- To evaluate the quality of your Gen-AI approach, you may do 2 things:
    1. Compare the histograms of columns D, E, G, H, and I of churn-gen-ai-test-data.csv and churn-bigml-20.csv datasets
    2. Test your trained model, as you did in Stage 4, with churn-gen-ai-test-data.csv and compare the classification results with the results of the original test dataset, churn-gen-ai-test-data.csv tests. You can use the classification accuracy, precision, and recall over here.
- It means that we expect your synthetic data to show rather the same statistical features as the original data set. Also, the trained models should show rather the same performance on both. You will not retrain the models, but just test them once again with your synthetic AI-Generated data.

6. **Model Interpretation and Insights:**
    - Identify key features influencing customer churn.
    - Provide actionable insights for customer retention strategies.
    - Determine the way Gen-AI can be employed in a data mining study.

7. **Reporting:**
    - Compile procedures, findings, visualizations, and conclusions into a comprehensive telling poster.

Tools and Libraries

The programming language is Python. Google Co-lab or Jupyter notebooks are advised. While coding is necessary, using any popular Python library is allowed. E.g.,

- Data Manipulation: Pandas, NumPy
- Data Visualization: Matplotlib, Seaborn

- Machine Learning Algorithms: Scikit-learn
- Gen-AI part: Any model and library that you need

Employing LLMs is fine for learning about models and algorithms. If that's the case, it should be mentioned in the poster. We don't advise direct usage of AI-generated codes.

Expected Outcomes

- Identification of distinct customer segments with similar behaviors.
- A predictive model capable of identifying customers at risk of churning.
- Insights into factors contributing to customer churn, aiding in the development of targeted retention strategies.

To Sum Up

1. This project offers a comprehensive experience in both supervised and unsupervised learning, providing students with practical skills in data preprocessing, modeling, and analysis within the context of a real-world application.
2. Please review the appendices, too.
3. In the evaluation process of your projects, we roughly consider the weights below.

Phase	**Score**
Poster Quality	3
Project trend and research method	6
Algorithms and methods used, unsupervised and supervised stages	5
Methods and results of the Gen-AI stage	5
Performances	4
Q and A	2
Total (CA Marks)	**25**

****

Appendix 1

More about the dataset

Each row represents a customer; each column contains a customer's attributes. The datasets have the following attributes or features:

- State: string
- Account length: integer
- Area code: integer
- International plan: string (binary Yes/No)
- Voice mail plan: string (binary Yes/No)
- Number vmail messages: integer
- Total day minutes: double
- Total day calls: integer
- Total day charge: double
- Total eve minutes: double
- Total eve calls: integer
- Total eve charge: double
- Total night minutes: double
- Total night calls: integer
- Total night charge: double
- Total intl minutes: double
- Total intl calls: integer
- Total intl charge: double
- Customer service calls: integer
- Churn: string (binary True/False)

The "churn-bigml-20" dataset contains 667 rows (i.e., customers, 2666 rows in the case of churn-bigml-80) and 20 columns (features).

The "Churn" column is the target to predict.

Appendix 2

About Your Poster

- To have some tips and templates regarding poster preparation
    - PosterPresentations.com: Offers a variety of free, professionally designed PowerPoint research poster templates in multiple sizes. These templates are fully customizable to fit your presentation needs.
    - Genigraphics: Provides free PowerPoint poster templates designed for scientific, medical, and research presentations. The templates are user-friendly and can help you achieve professional results.
    - GENIGRAPHICS PosterNerd: Offers free PowerPoint templates to assist in creating scientific posters. They provide various designs, including the "Billboard" style, aimed at simplifying information sharing.
    - Microsoft Create: Features customizable PowerPoint poster templates that are visually appealing and informative, suitable for various academic presentations.
    - Aresty Research Center at Rutgers University: Provides several research poster templates in PowerPoint format, along with guidelines for creating effective posters.
    - Springfield College Library Services: Offers PowerPoint poster templates and best practices for designing academic posters. Templates are available in various sizes and can be adjusted as needed.
- Which blobs/boxes we may see in your poster
    - Title and project general attributes
    - Declaration of team members and roles
    - Definition and goals of doing this project
    - Short description of the dataset
    - Preprocessing and its results (could be more than one box)
    - EDA (could be more than one box)
    - Unsupervised learning (could be more than one box)
    - Supervised learning (could be more than one box)
    - Gen-AI, one or two boxes
    - Conclusion, comprises the most important results, challenges you faced, and what you learnt.
- However, the poster format and size are up to the teams to determine.
- Some extra resources for academic posters:
    - https://our.umd.edu/poster-guide-design?utm_source=chatgpt.com

- https://researchguides.dartmouth.edu/c.php?g=1222845&p=8976597&utm_source=chatgpt.com

Appendix 3

Additional Resources

- *Towards data science* site can help you with the theory and background that you may need to carry out this mini-project, as well as the course slides and reference books.
- SKLearn documentations.
- KDnuggets website– Articles, tutorials, and industry insights on data mining, machine learning, and AI.
- DataCamp website – Interactive courses on data science, data mining, and Python/R for analytics.
- Machine Learning Mastery website– Guides and step-by-step tutorials on implementing data mining algorithms.