Project Midterm Report
Customer Retention Prediction Using Machine Learning

Team Name – Hive Mind

Team Members:

Dikshali Margaj - 801076473
Kshitij Shah - 801077782
Parth Mehta - 801057625
Upasana Pattnaik - 801081007

Introduction

A good method for companies to increase revenue is to invest in pre-existing customers. Retaining customers becomes essential for consumer dependent businesses. If customers are on the verge of leaving the service, it would be prudent for the company to identify what factors influence a customer's lack of interest in the product. As the cost of retaining a customer is far lesser than getting a new one, analyzing customer churn can reveal valuable insights. The loss of customers is known as customer churn or customer attrition. Companies analyze customer churn to uncover which factors lead to a customer voluntarily switching to a rival business.

Customer churn occurs when customers or subscribers stop doing business with a company or service. Also known as customer attrition, customer churn is a critical metric because it is less expensive to retain existing customers than it is to acquire new customers – earning business from new customers means working leads all the way through the sales funnel, utilizing your marketing and sales resources throughout the process. Customer retention, on the other hand, is generally more cost-effective, as you have already earned the trust and loyalty of existing customers.

Using a machine learning approach, we can find patterns which cause customer churn and forecast it to obtain a prognosis on which factors impact customer retention. It can help make sense of relationships between data. The models can predict customers with high probability to churn based on analyzing customers personal, demographic and behavioural data to provide personalized and customer-oriented marketing campaigns to gain customer satisfaction.

Problem Description

Currently, churn analysis of telecommunication dataset have different approaches to solving the problem of customer attrition. The focus is to find the key indicators of customer churn for our dataset and find a reliable metric to find the likelihood of an active customer leaving the company. We aim to contribute by using a different set of metrics to improve the results of regression and classification results.

Data

The data we will be using for our project is a Telecom Customer Churn. https://www.kaggle.com/blastchar/telco-customer-churn

The data set includes information about:

- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

The data set columns are as following**:**

- customerID: Unique Id for customers
- gender: Whether the customer is a male or a female
- SeniorCitizen: Whether the customer is a senior citizen or not (1, 0)
- Partner: Whether the customer has a partner or not (Yes, No)
- Dependents: Whether the customer has dependents or not (Yes, No)
- tenure: Number of months the customer has stayed with the company
- PhoneService: Whether the customer has a phone service or not (Yes, No)
- MultipleLines: Whether the customer has multiple lines or not (Yes, No, No phone service)
- InternetService: Customer's internet service provider (DSL, Fiber optic, No)
- OnlineSecurity: Whether the customer has online security or not (Yes, No, No internet service)
- OnlineBackup: Whether the customer has an online backup or not (Yes, No, No internet service)
- DeviceProtection: Whether the customer has device protection or not (Yes, No, No internet service)
- TechSupport: Whether the customer has tech support or not (Yes, No, No internet service)
- StreamingTV: Whether the customer has streaming TV or not (Yes, No, No internet service)
- StreamingMovies: Whether the customer has streaming movies or not (Yes, No, No internet service)
- Contract: The contract term of the customer (Month-to-month, One year, Two years)
- PaperlessBilling: Whether the customer has paperless billing or not (Yes, No)
- PaymentMethod: The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
- MonthlyCharges: The amount charged to the customer monthly
- TotalCharges: The total amount charged to the customer
- Churn: Whether the customer churned or not (Yes or No)

Literature Survey

1) The Analysis of Customer Churns in E-commerce based on Decision Tree [1]

Published in: 2015 International Conference on Computer Science and Applications (CSA)

Summary: The authors of this paper used the Decision Trees method to analyze an e-commerce's customer churn. They have approached this problem by using classification to predict customer churn. Using this method, they were able to find important features that attribute to attrition. They considered this method suitable for classification because of its straightforward approach. Their model has 88% prediction accuracy. They were able to discover that the discount revenue rate being less than 42.3% of average, classified the customer to be in sleep status. This provided us with an insight into how researchers approached customer churn.

Strengths and weaknesses: Decision tree model presentation and results are understandable by the stakeholders and is an intuitive approach to predicting customer churn. Their approach focuses on the the effect of using decision trees on their e-commerce dataset.

Relation to approach: We do not plan to limit ourselves to one model. Our approach will use decision trees to contrast its result with other machine learning models. Our focus is on the features present and developing relevant features to augment the dataset present and improve the results.

2) **Telecommunication Subscribers' Churn Prediction Model Using Machine Learning [2]**

Published in: Eighth International Conference on Digital Information Management (ICDIM 2013)

Summary: The researchers used Linear Regression, Logistic Regression, Artificial Neural Networks, K- Means Clustering and Decision Trees variations (CHAID, Exhaustive CHAID, CART, QUEST) to classify their 106,000-sample dataset from Customer DNA's website. Active and churn customers were classified using these models. In their literature survey covered paper which used Neural Networks and Regression models to predict customer churn. They faced class imbalance in their dataset, i.e., there were more active users than churn users. They used various re-sampling methods to handle this imbalance. Exhaustive CHAID model delivered the best results to predict customer churn.

Strengths and Weaknesses: Customer churn is exclusively a classification problem in this paper. The imbalance present in their dataset provided insight into handling datasets with skewed classes.

Relation to approach: This paper focused on classification models and their results on an e-commerce dataset. We plan to perform regression along with classification.

3) **Machine-Learning Techniques for Customer Retention: A Comparative Study [3]**

Published in: International Journal of Advanced Computer Science and Applications (IJACSA) 2018

Summary: The author's aim was to find a benchmark for churn classification using multiple machine learning model approaches. Comparing a host of models (Regression analysis: logistic regression, Decision tree–CART, Bayes algorithm: Naïve Bayesian, Support Vector Machine, Instance-based learning: k-nearest Neighbor, Ensemble learning: Ada Boost, Stochastic Gradient Boost and Random Forest, Artificial neural network: Multi-layer Perceptron, Linear Discriminant Analysis) on a telecommunication dataset with 3333 records. Evaluating each model's performance using model metrics, they have provided an in-depth analysis of how each model performs classification to find the customer churn.

Strengths and weaknesses: This comparative study provides us information on how each model performs on their low sample data.

Relation to approach: SVM gave the best results and could be considered as a contrast model in our approach to finding the best results for customer churn. They used K-fold cross-validation to present the results for each model. Our approach plans to focus on improving the feature set as opposed to a comparative study which is presented in this paper.

4) **Behavioural Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Custom Defection and Loyalty [4]**
Published in: 2015 IEEE International Congress on Big Data

Sumarry: The authors of this paper developed a "churn score" which they assigned to each customer to predict the early signs of customer churn to indicate the likelihood of customer churn. They used feature engineering to find the most relevant features, assign a probability(churn score) to each customer and subsequently input the features into different machine learning models.

Strengths and Weaknesses: This paper is closest to what we want to achieve from our project. Their feature engineering and feature selection presented an approach to a feature engineering focused analysis. They have a vastly different dataset with anonymous feature names, which they had to extract and build new features based on a supervised learning approach to finding a retention score based on probability.

Relation to approach: The dataset used is vast with over a billion samples and multiple features. They focus on subscription data of prepaid customers in a telecom dataset. It requires feature engineering and selection to find the best features to input into machine learning models. Our approach aims to use survival analysis to find the retention score of each customer as opposed to the inactive day's parameter used in this paper.

Method

Our new approach aims to find a different angle to work this problem. From our literature review, we were able to identify key processes that would add value to our project. The different papers provided us insight into how researchers approached customer retention with machine learning methods. Focusing on feature engineering, we hope to apply metrics like Customer Retention

Rate, develop a Retention Score and derive more features to add to our dataset. The retention score is based on survival analysis results.
Using Survival analysis for churn, we draw a parallel to the analysis being an expected survival statistic for biological organisms. It provides us a plot of survival against time. Survival regression model finds the effect of different features on survival over a period of time. This approach can aid in developing a probability based retention score based on a customer's location on the chart.
To elucidate further on Survival Analysis:
For example:
Consider 4 customer A, B, C, D
Cust A left after 3 months
Cust B left after 6 months
Cust C,D are still customers but we dont know whats the time to churn for them, to know this we will use survival model.

Survival model - 2 phase
1. Survival Analysis
2. Survival Regression

Survival Analysis(data visualization):
As used in medicine, it is used to check how long it takes before a person dies. Similarly, we can apply it to our case on how long it take before a customer churns.
We will use the survival curve which visualize the result over time. For example, how many years/months on average do customer stay, how long male customer stay compared to female customer or age group- to essentially understand the customer lifecycle.

Survival Regression: (applying the model on survival analysis)
We can show the probability of a customer and how long they would stay.
This can tell us what is driving the probability to churn, even for customers who did not churn yet. An example of a model is CoxPH model available in python.
Plan:

1- Data Visualization: The data preprocessing step is enhanced with visualizations to help us understand patterns in our data.

2- Data Preprocessing: Handling null values, categorical variables, scaling numerical values and dropping irrelevant columns.

3- Feature Engineering: This new approach focuses on feature engineering. Using survival analysis to understand the customer lifecycle we aim to develop a churn score. Customer retention score is the other metric derived by calculating the percentage of customer who has stayed over a given period and can be calculated on an annual, monthly or weekly basis. Deviating from the approaches encountered, we aim to add more features by including these metrics.

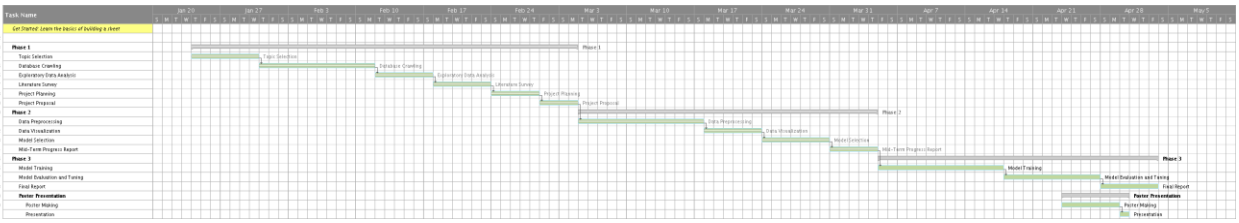
Customer Retention Rate = ((CE – CN) / CS)) x 100
CE = Number of customers at end of period CN = Number of new customers acquired during period CS = Number of customers at the start of a period

5- Analysis Setup: Weighing between oversampling and cost function-based approach to handle class imbalance, we will be choosing our approach to address our class imbalance problem present in our dataset. We feed the augmented dataset into different machine learning models and compare the results.

i) K-Means Clustering: Unsupervised learning approach to find relations between different features.
ii) Logistic Regression: Classification approach to customer churn.
iii) K-Nearest Neighbor: Classification approach to customer churn.
iv) Decision Trees: Feature importance indicator
v) Neural Networks: Non-linear approach to classification
vi) Survival Regression: Survival analysis-based prediction model

Difficulties Faced During Implementation

Timeline



For better visualization of the timeline please refer to the below hyperlink.

Timeline

Accomplished Milestones

- Topic Selection
- Dataset Crawling
- Exploratory Data Analysis
- Literature Survey
- Project Planning
- Project Proposal
- Data Preprocessing
- Data Visualization

Work assignment

Task	Dikshali Margaj	Kshitij Shah	Parth Mehta	Upasana Pattnaik
Topic Selection	✓	✓	✓	✓
Dataset Crawling	✓	✓	✓	✓
Exploratory Data Analysis	✓	✓	✓	✓
Literature Survey	✓			✓
Project Planning	✓	✓	✓	✓
Project Proposal	✓	✓	✓	✓
Data Preprocessing	✓	✓	✓	✓
Data Visualization		✓	✓	
Model Selection	✓	✓	✓	✓
Mid-Term Progress Report	✓	✓	✓	✓
Model Training	✓	✓	✓	✓
Model Evaluation and Tuning	✓	✓	✓	✓
Final Report	✓	✓	✓	✓
Poster Making	✓	✓	✓	✓
Presentation	✓	✓	✓	✓

Is our idea novel?

All subscription-based services have been trying to figure out ways to retain customers. We will be focusing on the feature engineering and the class imbalance present in the dataset. Looking to get more insight of the dataset. Applying a set of metrics that we hope can add a difference to the problem of the customer churn. We are performing clustering technique to segment our data. The different segementation categories are like best loyal customers, almost best lost customers and lost cheap customers.

Response to feedback

We focused the literature survey on the relevance of our approach with respect to the methodologies used by the authors and described the limitations and their strengths.
Addressing the lack of difference in our approach, we dug deeper into different kinds of literature present on customer churn. We noticed that feature engineering of the telecom dataset was sparse and the metrics that we decided to use were not present. We drew our attention to finding new features to augment our analysis.
We have added our preprocessing steps in our methodology.
We updated the timeline structure to with subtasks and a proper Gannt chart of the timeline.
We have updated our work distribution to be more detailed.
Our new approach to the project's novelty differs from the original aim by focusing on feature engineering as opposed to find relevant features. We are working on adding new features to the data based on metrics we are developing to augment our analysis.

References

[1] F. Guo and H. Qin, "The Analysis of Customer Churns in e-Commerce Based on Decision Tree," 2015 International Conference on Computer Science and Applications (CSA), Wuhan, 2015, pp. 199-203.
doi: 10.1109/CSA.2015.74
keywords: {competitive intelligence; customer relationship management; data mining; decision trees; electronic commerce; customer churns; e-commerce; decision tree; competitiveness; customer relationship management; business decisions; data mining; business customers; Decision trees; Companies; Training; Customer relationship management; Electronic commerce; Probability; e-commerce; decision tree algorithm; customer churns; model},
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7810863&isnumber=7810811

[2] S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal and A. Rehman, "Telecommunication subscribers' churn prediction model using machine learning," Eighth International Conference on Digital Information Management (ICDIM 2013), Islamabad, 2013, pp. 131-136. doi: 10.1109/ICDIM.2013.6693977 keywords: {competitive intelligence;customer profiles;data mining;decision trees;learning (artificial intelligence);mobile communication;neural nets;pattern classification;regression analysis;telecommunication services;telecommunication subscriber churn prediction model;machine learning;market saturation;public policies;mobile communication standardization;customer switching;fluid market;mobile carriers;customer acquisition;customer retention;business intelligence application;customer identification;data mining technique;historical data;pattern finding;regression analysis;decision trees;artificial neural networks;ANN;customer DNA Website;traffic data;usage behavior;resampling method;class imbalance;classifier algorithm;potential churner identification;Decision trees;Prediction algorithms;Logistics;Predictive models;Correlation;Linear regression;Churn prediction;Business Intelligence;Data Mining}, URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6693977&isnumber=6693962

[3] Sahar F. Sabbeh " Machine-Learning Techniques for Customer Retention: A Comparative Study". International Journal of Advanced Computer Science and Applications, Vol. 9, No. 2, 2018

[4] M. R. Khan, J. Manoj, A. Singh and J. Blumenstock, "Behavioral Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Custom Defection and Loyalty," *2015 IEEE International Congress on Big Data*, New York, NY, 2015, pp. 677-680.
doi: 10.1109/BigDataCongress.2015.107
keywords: {consumer behaviour;feature selection;learning (artificial intelligence);customer identification;mobile phone network;supervised learning algorithms;feature selection;brute force approach;churn score;custom defection;churn prediction;behavioral modeling;Measurement;Accuracy;Predictive models;Prediction algorithms;Mobile handsets;Mobile communication;Supervised learning;Churn;machine learning;supervised learning;data science;call detail records},
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7207291&isnumber=7207183

Github repo address

https://github.com/shahksh1011/Customer-Retention-Predictive-model