**Churn Risk Score**

**Team Details.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Email | Country | College/Company | Specialization |
| Fabian Umeh | Fabianumeh335@gmail.com | UK | Teesside University | Data Science |
| Rukevwe Ovuowo | rukevwe10@gmail.com | Nigeria | GBG Data science Academy | Data Science |
| Olutayo Oladeinbo | oladeinboolutayo@yahoo.com | UK | Teesside University | Data Science |

**Abstract**

Churn rate is a metric used to determine how many clients or staff leave a business in a certain time frame. It could also refer to the sum of money that was lost because of the departures. The goal of this project is to develop and execute a model that predicts how long a person will remain a client by focusing on their browsing habits and previous purchase information, among other details. Every day, this rating is updated for all users who have had at least one interaction. The allocated value ranges from 1 to 5. Numerous strategies were used throughout the model's construction to prepare the dataset for the machine learning and deep learning models to produce accurate predictions. The dataset was processed using methods such managing missing data points, duplication checks, feature engineering, date formatting, and data encoding. Algorithms like Random Forest, KNN, Multi-layer Perceptron, and Deep Learning Models were employed in the training and testing phases, with the deep learning model performing better compared to the rest.

**Introduction**

**What is Churn rate?**

The rate at which consumers cease doing business with a company is known as the churn rate, sometimes referred to as the rate of attrition or customer churn. The most popular way to represent it is as the proportion of service users that cancel their memberships within a predetermined time frame. It is also the frequency of employees quitting their positions within a predetermined time frame. A company's growth rate (determined by the number of new customers) must be higher than its attrition rate to grow its customer base.

The profitability and growth might be negatively impacted by a high churn rate. The churn rate is a crucial element in the telecom sector. Since many of these businesses are in competition with one another, switching service providers is usually simple for consumers.

The churn rate considers both moving providers and discontinuing service without switching carriers. In subscriber-based firms where subscription payments make up most of the income, this metric is particularly useful. Depending on the industry, a turnover rate may be deemed positive or negative.

**Pros and cons of churn rate**

Pros

* Provides clarity on the quality of the business
* Indicates whether customers are satisfied or dissatisfied with the product or service
* Allows for comparison with competitors to gauge an acceptable level of churn
* Easy to calculate

Cons

* Does not provide clarity on the types of customers leaving; new versus old.
* Does not differentiate between types of companies in industry comparison; start-ups, growing, and mature.

Methodology

Data Exploration and feature selection

The data used for this project was gotten from hackerearth churn competition. The dataset has 25 features and 36992 data points. Features that were not useful to the analysis were removed they include customer\_id, referral\_id, security\_no and name. all features were of data type; object except the average time spent, average transaction value, points in wallet, days since last visit, churn risk score and age.

Data Cleaning

Below are the list of problems found in the dataset and the respective approaches used in handling them.

1. **Missing values:**

Some values in columns such as [region\_category, preferred\_offer\_types, points\_in\_wallet], appears to be missing.

**Approach:**

* The region category was encoded as follows:

City – 3

Town – 2

Village – 1

And as such, missing values were filled with the non-extreme value (2) for town.

* The points\_in\_wallet column had a relatively small percentage of missing values, and these values were filled using the average of the collected samples.

* The Preferred offer type column was encoded as follows:

Offer – 1

Without offer – 0

And as such the few missing values were replaced with no offers.

1. **String Error:**

The data type in the average frequency login days column was supposed to be ‘int’ but rather contained some string value ‘Error’, which was further removed.

**Approach:**

The average frequency login days column keeps record of the average days the company site is visited by a customer. So ideally ‘Error’, most likely represent 0 login days was replaced as such.

1. **Negative values:**

Some columns contained negative values, which appeared to be anomalous and were further excluded from the analysis.

**Approach:**

The negative (anomalous) values were few and were excluded from the analysis.

Data Transformations

Data transformation are thevarious formatting applied to the dataset to make it easy for the algorithm to decipher. Below are a list of transformations performed on the dataset.

1. Removing Unused columns

Some features appeared to be irrelevant in the prediction of customers churn score rate, and as such were excluded from the study.

1. Feature Engineering

The data in its raw form can only provide limited information. Some transformation techniques were performed on the data, this is to convert the data into a machine-readable format. Some of the transformation includes: date formatting, binary encoding and categorical encoding.

* Date Formatting
* Getting the number of days spent by each customer
* Getting the number of hours spent in the last visit by each customer
* Low-level categorical feature format

Some categorical features has some classes either few or unknown. To handle this, the said classes were split equally between the two major class as shown below.

* Generalizing categories to reduce complexity

Classes that represent the same phenomena were further classified as one, for instance the feedback columns with categories such as ‘Poor Product Quality’, ‘Poor Website’, ‘Too many ads’, ‘Poor Customer Service’ all represent bad reviews.

* Binary Encoding
* Categorical encoding

1. Finally, Handling of missing values

Result

The target variable is churn risk score and the evaluation metric used as per highlighted by the task is stated below.

score **=** 100 **\*** f1\_score(test, predictions, average**=**"macro")

The results of the machine learning models are summarized below

|  |  |  |
| --- | --- | --- |
| Model | Model parameters | score |
| KNN | n\_neighbors=14 | 46.00 |
| MLP | hidden\_layer\_sizes **=** 8, solver**=** 'lbfgs' | 51.72 |
| DL | Dense layer (16, activation = ‘relu’)  Dense layer (8, activation = ‘relu’)  Dense layer (6, activation = ‘sigmoid) | 70.13 |

**Final Recommendation**

Personally, I think the way the data was aggregated makes it difficult to identify which range of score depicts churning or not churning for instance 3 and above may represent churn and below not churn.

If simplified into churn or not churn, this will help the model generalise better. Though it still depends on the usage of the model.