PROJECT REPORT

ON

COSTUMER CHURN ANALYSIS

BY

EMMANUEL UNGWADA

INTRODUCTION

The problem is based on when the company's customers stop doing business with the company. Businesses are anxious on measuring agitation because keeping existing customer is far less expensive than getting a new customer. New enterprise includes operating leads through an sales funnel, the usage of advertising and marketing and income budgets to advantage extra clients. Existing clients will frequently have a better quantity of provider intake and might generate extra client referrals.

Customers maintaining is the maximum essential asset for any enterprise as it's far said that "the fee of obtaining a brand new purchaser may be better than that of maintaining a purchaser via way of means of as tons as 700%; growing purchaser retention costs via way of means of a trifling 5% ought to growth income via way of means of 25% to 95%". So one of the high-quality way to keep the clients is to lessen churn rate, wherein "churn" approach shifting the purchaser from carrier company to every other one, or preventing the use of precise offerings over precise duration for plenty motives that may be detected formerly if the employer analyzes its statistics facts and makes use of gadget mastering generation which permits the organizations to expect the clients who're probable to churn. A lot of research authorized its performance to this situation so the employer can reply speedy to the behavioral modifications in the purchaser’s minds. Businesses nowadays is refining & optimizing the purchaser revel in that's the important thing to maintaining a marketplace differentiation and lowering churn [5], wherein maintaining an present purchaser expenses tons decrease than obtaining a brand new one. This studies research the gadget mastering algorithms and advocated the high-quality answers for business. In the aggressive business sector, clients can without difficulty transfer from one company to every other, which we could the business vendors concerned approximately their clients and a way to keep them however they are able to expect the clients who will pass to every other company formerly via way of means of studying their behavior. They can keep them via way of means of supplying gives and their favored offerings in step with their ancient facts so the goal of this take a look at is to expect churn formerly and come across the principle elements that could permit the consumer pass to every other company in business.

Work related

Many research are available for churn problem from one-of-a-kind viewpoints with one-of-a-kind datasets, set of rules and for one-of-a-kind industries in which churn evaluation is one of the international extensive used to investigate the consumer behaviors and expect the clients who're approximately to go away the carrier settlement from a corporation. Studies discovered that gaining new clients is five to ten instances more expensive than preserving current clients glad and dependable in today ’ s aggressive conditions, and that a mean corporation loses 10 to 30 percentage of clients annually [6] [7]. Most of the literature targeted extra on information mining algorithms, however only some of them targeted on distinguishing the critical enter variables for churn prediction and on improving the information samples thru green pre-processing for use for information mining algorithms implementation [8] [9]. Amin, A., et al. [10] offered a singular churn prediction technique primarily based totally at the classifier’s reality estimation the use of distance aspect in which they grouped the dataset into one-of-a-kind zones primarily based totally on the gap which can be then divided into classes with excessive and occasional reality, they used four datasets with one-of-a-kind samples and that they had been discreet through size, the values that exists in every characteristic of the dataset, after which assigned positive labels and on the give up produced precise listing of values in one-of-a-kind range of businesses of an characteristic.

What is the problem?

The data set contains 7043 rows and 21 columns. The main problem is Customer retention may be executed with properly customer support and products. But the handiest manner for a agency to save you attrition of clients is to simply understand them. The giant volumes of facts amassed approximately clients may be used to construct churn prediction models. Knowing who's maximum probably to disorder approach that a agency can prioritise targeted advertising efforts on that subset in their client base. Preventing client churn is seriously vital to the telecommunications sector, because the limitations to access for switching offerings are so low.

This report focuses on customer data from IBM sample Data sets with the aim of building and comparing several customer churn prediction models.

DATA ANALYSIS

The fundamental method of fixing this problem was first reading the data , then bringing out insights from the dataset and after that I even have accompanied a device mastering pipeline a good way to remedy the problem.

The machine learning conveyor that is used include;

- importing libraries that Is necessay for the dataset
- performing data preprocessing
- modelling using logistic regression
- performing prediction

The domain used was jupyter notebook and the libraries exploited include numpy, matplotlib, KNN classifier, pandas, seaborn, plotly.graph_objects were use for computations.

TASK DEFINITION

- KNN Classifier: K Nearest Neighbor(KNN) is a completely simple, clean to understand, flexible and one of the topmost system mastering algorithms. KNN used withinside the sort of packages which includes finance, healthcare, political science, handwriting detection, photo reputation and video reputation. In Credit ratings, economic institutes will are expecting the credit score score of customers. In mortgage disbursement, banking institutes will are expecting whether or not the mortgage is secure or risky. In political science, classifying capacity citizens in instructions will vote or won't vote. KNN set of rules used for each category and regression problems. KNN set of rules primarily based totally on function similarity approach.

- Logistic Regression:

Classification strategies are an important a part of system mastering and records mining applications. Approximately 70% of issues in Data Science are category issues. There are plenty of category issues which might be available, however the logistics regression is not unusualplace and is a beneficial regression approach for fixing the binary category problem. Another class of category is Multinomial category, which handles the problems in which more than one lessons are gift withinside the goal variable. For instance, IRIS dataset a completely well-known instance of multi-elegance category. Other examples are classifying article/blog/report class. Logistic Regression may be used for diverse category issues including junk mail detection. Diabetes prediction, if a given purchaser will buy a selected product or will they churn any other competitor, whether or not the consumer will click on on a given commercial hyperlink or not, and lots of greater examples are withinside the bucket. Logistic Regression is one of the maximum easy and typically used Machine Learning algorithms for two-elegance category. It is straightforward to enforce and may be used because the baseline for any binary category problem. Its simple essential ideas also are positive in deep mastering. Logistic regression describes and estimates the connection among one established binary variable and unbiased variables.

DATASET

The below subset of the data was extracted from IBM which include;

- Customer ID
- Gender
- SeniorCitizen
- Partner
- Dependent
- Tenure
- Phoneservice

Data processing

In order to become aware of the function which can be maximum predictive of churn rates, it's miles first important to outline what variable can be used because the goal occasion this is the ‘churn’ being predicted. In this case, I pick out to apply the Page variable because the churn occasion, and subset the facts to encode a price of ‘1’ whilst an present consumer both downgrades a subscriptions of confirms the cancellation of subscription, and encoded a price of ‘0’ otherwise.

Looking at the level variable, 79% of users are currently paying for the service while 20.8% are currently using it for free

Looking on the correlation among variables, there does now no longer look like anyone variable this is quite correlated with another, other than the connection among a consultation ID and the period of a consultation (despite the fact that this has now no longer inherent meaning)

Model implementation and refinement

Once the datasets have been balanced and Exploratory Data Analysis performed, I created Indexes from the Churn, Gender, Level, Location, and User Agent fields and in shape them right into a ordinary distribution in order that they will be dealt with as factorized columns to be inputted right into a system getting to know primarily based totally version. For the very last a part of the records preparation, I introduced columns for the range of instances the Save Settings Page became accessed, range of songs a consumer listened to, counts of Thumbs-up and Thumbs-down rankings, the range of instances a consumer Added a tune to a playlist, and the common remember of songs consistent with session. The very last columns for use in the version are: gender_index, level_index, user Agent_index, ts_day, saved_settings, thumbs_up, thumbs_down, playlist_added, and songs_per_session . With the records organized to be fed into the version, I used a String Indexer to forged the 'churn' column right into a column of label indices, a Vector Assembler to mix all the columns right into a unmarried Vector Column, after which use a Normalizer to normalize the ensuing vector alongside a wellknown distribution. The dataset became then break up right into a 80% schooling and 20% checking out partition to be feed into the version, and a Multiclass Evaluator became created to study the consequences after they may be run via the version. Four Modeling Techniques have been then carried out to expect churn rates: a Logistic Regression Model, a Random Forest Classifier, a Gradient Boost Tree Classifier.

- The Logistic Regression version will degree the enter functions chance of as it should be expect a churned customers primarily based totally on a scale starting from zero to 1, with zero representing no chance of an correct prediction and 1 representing a superbly correct prediction.

- The Random Forest Classifier will use selection bushes to carry out iterative operations comparing the fine aggregate of enter functions to efficaciously expect consumer churn rates.

-The Gradient Boost Tree Classifier makes use of comparable selection tree approaches, besides on smaller subsets of the records after which the usage of averaging strategies to pick out the maximum finest aggregate of bushes to maximise the chance of an correct prediction.

Model evaluation

Two capabilities had been created to assess the end result sets: an Evaluation Function to output the F1 score, Weighted Precision, Weighted Recall, and Accuracy Rates, and a Confusion Matrix to output the True Positive, True Negative, False Positive, False Negative, Precision and Recall rates.

Model improvement

This model can be progressed via way of means of the use of unique sampling strategies to balance the authentic dataset, such as random over or below sampling, artificial over or below sampling, and a few mixture of the 2 approaches. More approximately sampling techniques may be examine here. When acting Exploratory Data Analysis at the authentic dataset, a few lacking values had been found for positive columns, and the facts with lacking values had been eliminated from the information sets. This might also additionally have negatively impacted the accuracy of the models, because the eliminated facts might also additionally have had a significant courting with the churn rate. Instead of casting off the facts with null values, I ought to have used substitution strategies inclusive of sampling with replacement, sampling with out replacement, or stratified sampling to boom the accuracy of the version. Alternatively, Principle Component Analysis might have been used to decide the maximum most suitable set of capabilities in the enter dataset the use of capabilities inclusive of eigenvalues to decide which enter variables seize the most quantity of variance in the churn rates, then those might have been used to populate the authentic models. Larger capabilities ought to have additionally been damaged up into smaller capabilities, such the use of regex to interrupt up the ‘location’ column into unmarried city-nation values or breaking the ‘userAgent’ column up into a couple of columns for the customers browser and running system. Finally, after strolling our Gradient Boost Tree Model, hyper parameters might have been used to music the version’s Pipeline and Parameter Grid values to discover the most suitable set of attributes to music the version via way of means of minimizing a loss function.

Reference:

Andrews, R., et al. (2019) Churn Prediction in Telecom Sector Using Machine Learning. International Journal of Information Systems and Computer Sciences, 8, 132-134. https://doi.org/10.30534/ijiscs/2019/31822019

ApurvaSree, G., et al. (2019) Churn Prediction in Telecom Using Classification Algorithms. International Journal of Scientific Research and Engineering Development, 5, 19-28.

Tata Tele Business Services (2018) Big Data and the Telecom Industry.

Kayaalp, F. (2017) Review of Customer Churn Analysis Studies in Telecommunications Industry. Karaelmas Science Engineering Journal, 7, 696-705.