Customer Churn Analysis

Big Data-BIA 678 Term Paper

Submitted By:
Savleen Kaur
10476867

INTRODUCTION

Customer turnover is a critical issue that is frequently linked to the business's current cycle. Sales expand at an exponential rate throughout the growth phase of a company's life cycle, and the number of new clients significantly surpasses the number of churners. Firms in a stable stage of their life cycle, on the other hand, place a high priority on minimizing client churn rates. A customer Churn rate is described as the annual percentage rate at which customers stop using a service. The most common causes of client churn are split into two categories: unintentional and purposeful. Accidental churn occurs when circumstances change, preventing customers from using the services later. For example, financial situations may make advantages too expensive for the customer. Intentional (also known as Voluntary) churn occurs when customers switch to another company that provides comparable services, such as better ideas from competitors, more developed services, and a lower cost for a similar service. In any organization, it's critical to understand the reasons that lead to voluntary client churn. Various market areas have some common aspects, such as how intertwined the client is with the firm.

Churn prediction has been a major concern in all the business' in recent years. To address this issue, for example, telecom carriers identify these clients before they leave. As a result, it's critical to create a unique classifier that can forecast future churns. To enhance revenue, three key tactics have been proposed: (1) attract new consumers, (2) upsell existing customers, and (3) extend client retention periods. However, when comparing these tactics based on the value of return on investment (RoI), it was shown that the third technique is the most profitable, demonstrating that retaining an existing customer is far less expensive than obtaining a new one, as well as being much easier than upselling. To implement the third method, businesses must reduce the risk of customer churn, sometimes known as "moving from one provider to another." Companies must be able to identify users who are likely to churn in the near future, allowing the operator to respond quickly with relevant discounts and promotions. Learning algorithms for classification, such as decision trees, logistical regression, k-nearest neighbors, Nave Bayes, neural networks, and others, are the most commonly used techniques for this purpose. Furthermore, researchers concentrate on uncovering new traits that are the most successful at forecasting customer attrition.

Using Telco customer data, I investigated the primary reasons for churn among consumers in this research. I acquired and processed data for this purpose, then implemented 3 algorithms based on it.

Challenges in Customer Churn

The purpose of customer churn prediction is to predict impending churners using data linked with each network user and a predetermined forecast horizon. The customer churn prediction problem is typically divided into three steps, according to Umayaparvathi & Iyakutti: training, test, and prediction. Previous data, like call records and personal and client's data, obtained and maintained by telecommunications service providers, contributes to the customer turnover problem throughout the training period. Furthermore, during the training stage, the labels are organized in the list of churners' recordings. In the test stage, the trained model with the highest accuracy is assessed to predict the churners' records from actual data. Customer churn prediction helps customer relationship management (CRM) avoid losing customers by advising retention strategies and improved incentives or packages to retain existing customers. As a result, a possible revenue loss for the company can be avoided.

Involuntary and voluntary churners are the two types of churners. Involuntary churners are customers who are dropped by their telecommunications service provider due to non-payment, deception, or non-use of the phone. Voluntary churners are customers who opt to quit their service with their telecoms service provider. Such churn is caused by a variety of variables, including economic 5 elements (for example, price sensitivity), technical considerations (for example, another telecommunications service provider offers more innovative technology), bad customer service factors, and other inconvenience-related factors. Identifying involuntary churners are simple; however, identifying voluntary churners is more complex. In general, the telecoms industry's customer churn problem is voluntary.

Algorithms for Customer Churn

Only a few studies have focused on finding the important input characteristics for churn prediction to be used for data mining methods implementation. Logistic Regression, Bayes Network, decision tree, and GBT Classifier, among other data mining techniques, were determined to be the most common algorithms in customer churn prediction. Some academics have also integrated a few algorithms to create an innovative algorithm with a higher accuracy rate.

I chose logistic regression model because it will identify relationships between our target feature, Churn, and our remaining features to apply probabilistic calculations for determining which class the customer belongs to.
I have also chosen Random Forest as it results in better predictive outputs.
Gradient Boosted Tree (GBT) is used for modeling. It is also used for ranking in the field of learning.

Data Collection and Preparation

Dataset- https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113 , https://community.ibm.com/community/user/datascience/communities/community-home/digestviewer/viewthread?GroupId=5059&MessageKey=5423f5a3-371d-4b05-aecf-7802d4f0b030&CommunityKey=d388e552-3a57-460d-911b-8da3b9e1e497&tab=digestviewer

Customer master, enrollments, and billing tables are the three portions of the specified data sets. To further cleanse, process, and generate a new feature from the existing characteristics, a data extract from these tables is obtained. The data set contains information about 7055 records in total with 20 features, and also has the following information:

- Churn- Customers who left within the last month
- User ID: Primary Key
- Days Active
- Settings
- Services for every individual customer– phone, multiple lines, internet etc
- Time- tenure being a customer
- Payment ways
- Monthly charges
- Total charges

Because we employed feature tools to create a large number of new attributes, we'll face a new problem: deciding which features to consider. To begin, we'll fit a traditional decision tree and then extract the feature that divides a big quantity of data amongst estimators. We can aid feature selection by obtaining the top factors that explain the majority of the variations.

```
# Split data into train and test set
events_pivot = events_pivot.withColumnRenamed('Churn', 'label')
training, test = events_pivot.randomSplit([0.7, 0.3])

# Create vector from feature data
feature_names = events_pivot.drop('label', 'userId').schema.names
vec_asembler = VectorAssembler(inputCols = feature_names,
                               outputCol = "Features")

# Scale each column
scalar = MinMaxScaler(inputCol="Features",
                      outputCol="ScaledFeatures")

# Build classifiers
rf = RandomForestClassifier(featuresCol="ScaledFeatures",
                            labelCol="label",
                            numTrees = 50,
                            featureSubsetStrategy='sqrt')
lr = LogisticRegression(featuresCol="ScaledFeatures",
                        labelCol="label",
                        maxIter=10,
                        regParam=0.01)
gbt = GBTClassifier(featuresCol="ScaledFeatures",
                    labelCol="label")

# Consturct 3 pipelines
pipeline_rf = Pipeline(stages=[vec_asembler, scalar, rf])
pipeline_lr = Pipeline(stages=[vec_asembler, scalar, lr])
pipeline_gbt = Pipeline(stages=[vec_asembler, scalar, gbt])

# Fit the models
rf_model = pipeline_rf.fit(training)
lr_model = pipeline_lr.fit(training)
gbt_model = pipeline_gbt.fit(training)
```

We undertake some further data analysis once we construct the underlying data collection to better understand the data, such as churn by gender, service type, time duration, etc.

Model Evaluation

Fitted models are placed to test first, and then the best performing one is selected.

```
def modelEvaluations(model, metric, data):
    """ Evaluate ML model's performance
        Input:
            model - pipeline object
            metric - the metric of the evaluations
            data - data being evaluated
        Output:
            [score, confusion matrix]"""

    # generate predictions
    evaluator = MulticlassClassificationEvaluator(
            metricName = metric)
    predictions = model.transform(data)

    # calculate score
    score = evaluator.evaluate(predictions)
    confusion_matrix = (predictions.groupby("label")
                        .pivot("prediction")
                        .count()
                        .toPandas())
    return [score, confusion_matrix]

f1_rf, conf_mtx_rf = modelEvaluations(rf_model, 'f1', test)
f1_lr, conf_mtx_lr = modelEvaluations(lr_model, 'f1', test)
f1_gbt, conf_mtx_gbt = modelEvaluations(gbt_model, 'f1', test)
```

The F1 score for the random forest model: 0.7777777777777779

	label	0.0	1.0
0	0	32	3.0
1	1	5	NaN

The F1 score for the logistic regression model: 0.7640845070422536

	label	0.0	1.0
0	0	31	4.0
1	1	5	NaN

The F1 score for the gradient boosting model: 0.803846153846154

	label	0.0	1.0
0	0	28	7
1	1	2	3

As per the F1 score, GBT model outperformed other two ie RF and LR models.

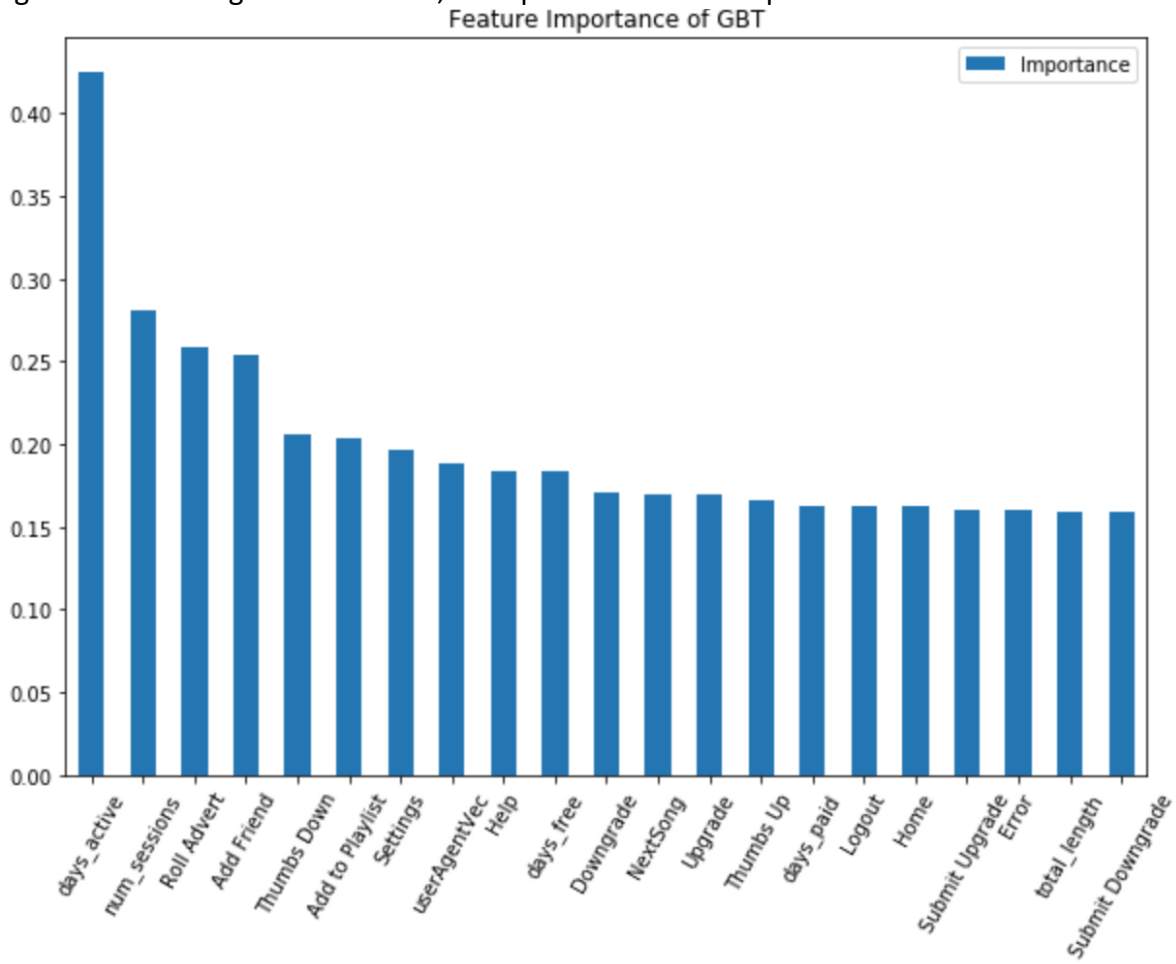
We'd rather our system correctly identify churning consumers than focus on overall performance because a small fraction of users will definitely churn. Let's suppose this: even if just 6% of consumers churn within the true population distributions, predicting everyone as non-churn gives us a 94% accuracy. Poor performance in a particular class, on the other hand, will be penalized by the F1 score, effectively reducing such issues. The imbalance is present in every churn forecast situation.

Feature Importance

The relative relevance rank of each feature we constructed will be visualized using the feature importance function. We'll treat userAgentVec as an encoded vector because it's the last characteristic. In the one-hot encoded vector, the code below adds up all of the feature significance values for all of the sub-features.

```
feature_importances = np.array(gbt_model.stages[-1]
                    .featureImportances)userAgentVec = feature_importances[len(feature_names)
:].sum()
feature_importances = feature_importances[:len(feature_names)]
                + [userAgentVec]
```

For gradient boosting trees classifier, we'll plot the feature importance.



The majority features we developed are significant contributors to user churning, with days_active being the most relevant.

Given that we know the number of active days is the most essential element, we may urge senior authorities to build an incentive system to encourage low activity users to stay online for longer periods of time.
Furthermore, because the agents that users used to access the service are so critical, we could pinpoint the falling agent and the engineering team starts to work on it specifically to resolve the problem.

How to Reduce Customer Churn

- Concentrate efforts on the most valuable consumers: Rather than focusing solely on providing incentives to customers who are on the verge of leaving, it may be more effective to invest your resources in your most loyal and profitable clients.

- Examine churn as it happens: Make use of your churned customers to figure out why they're going. Analyze how and when a client churns during their stay with your organization, and utilize that information to implement preventative actions.

- Demonstrate that you care about your customers: Trying a more proactive approach to connecting with your customers rather than waiting for them to contact you. Communicate all of the benefits you provide and demonstrate that you care about their experience, and they'll be more likely to stay. Social media can be utilized wisely to convey the benefits.

Advantages

- Examining a company's churn rate reflects the quality of the service it provides and its usefulness. It also shows if the client relationship is maintained.

- When a company's churn rate increases over time, it acknowledges that a key component of its business model is flawed. The company may have a malfunctioning product, poor customer service, or a product that isn't enticing to individuals who believe the price isn't worth the benefit.

- A high rate indicates that the company must figure out why customers are leaving and how to address the problem. The cost of getting new consumers is more than the cost of retaining existing customers, this leads to assessing the quality of the business while attempting to ensure that the clients you worked so hard to get remain paying customers.

Disadvantages

- It does not consider the kind of customers who are leaving. Customers who have recently joined show more signs of degradation.

- It is critical to think about the consequences of losing new customers versus long-term customers. Existing customers must have a solid reason to change their loyalty, however new customers are ephemeral. A high churn rate in one quarter doesn’t show product’s quality, it may represent a high growth rate in the previous quarter.

- Perhaps the business recently ran a campaign that drew in new clients. Customers who were trying out the product may decide it's not for them and cancel their membership once the campaign expired or even if the benefit of the deal never ended.

- Churn rate does not provide a meaningful industry comparison of the different sorts of businesses within a given industry. As new customers try out the business, most new businesses will have a high acquisition rate, but they will also have a high churn rate as these new clients leave.

CONCLUSION

The study was to understand the reason behind customer churn in telecom market and to assist businesses in retaining the profits. Predicting churn is a critical source of continuous revenue for telecom industry. Random Forest, Logistic Regression and GBT Classifier are examples of ML models that can help determine which clients are likely to leave and how to best service them. According to our findings, the GBT Classifier should be employed since we want our model to correctly identify the fundamental cause of churn, which in this case is Active Days. Companies can look forward to improving their policies including the pricing, schemes etc as they impact the active users in a day. The organization can implement following with this information:

- Relationship Managers (RM) get everyday updates about who is likely to churn with the reason for it.
- Customer grievances are noted and timely resolutions are provided, preventing the customer churn.
- Customer Engagement is enhanced when the RM reaches out to the customers, resolving their grievance, hence targeting customer emotions to increase the trust for their product.
- Operational costs of call centers can be reduced by dealing with the customer churn and dispute calls.

REFERENCES

https://ieeexplore.ieee.org/document/7684171
https://towardsdatascience.com/understanding-customer-churning-with-big-data-analytics-70ce4eb17669
https://doi.org/10.1177/0092070300281014
https://www.hindawi.com/journals/ddns/2021/7160527/
https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6

https://www.researchgate.net/publication/348148507_ECommerce_Customer_Churn_Prediction_By_Gradient_Boosted_Trees