Machine Learning Engineer Nanodegree

Capstone Proposal

Title: Smart Inventory

Domain Background

Most telecom companies suffer from voluntary churn. Churn rate has strong impact on the life time value of the customer because it affects the length of service and the future revenue of the company. For example if a company has 25% churn rate then the average customer lifetime is 4 years; similarly a company with a churn rate of 50%, has an average customer lifetime of 2 years. It is estimated that 75 percent of the 17 to 20 million subscribers signing up with a new wireless carrier every year are coming from another wireless provider, which means they are churners. Telecom companies spend hundreds of dollars to acquire a new customer and when that customer leaves, the company not only loses the future revenue from that customer but also the resources spend to acquire that customer. Churn erodes profitability.

Churn (loss of customers to competition) is a problem for telecom companies because it is more expensive to acquire a new customer than to keep your existing one from leaving.

In the targeted approach the company tries to identify in advance customers who are likely to churn. The company then targets those customers with special programs or incentives. This approach can bring in huge loss for a company, if churn predictions are inaccurate, because then firms are wasting incentive money on customers who would have stayed anyway. There are numerous predictive modeling techniques for predicting customer churn. These vary in terms of statistical technique (e.g., neural nets versus logistic regression versus survival analysis), and variable selection method (e.g., theory versus stepwise selection).

Check out the interesting article on customer churn from https://ieeexplore.ieee.org/document/8258400/

There are much more interesting articles related to “Business model and research on telecommunication operation transformation available at https://ieeexplore.ieee.org/document/5705218/?reload=true

Also check out https://data.gov.in where we have further research done in this area .you can obtain datasets related to Trends of Telecom subscribers in India as well as other areas e.g Telecommunication performance,Trends of Telecom subscribers in India, Internet and BroadBand subscribers ,teledensity in India etc

Problem Statement

This Capstone project is about enabling churn reduction using analytics and then implementing the most optimal Machine Learning model to reduce churn.

Datasets and Inputs

The datasets are taken from

- **https://www.kaggle.com/becksddf/churn-in-telecoms-dataset/data**
- **https://www.crowdanalytix.com/contests/why-customer-churn/**

The objective of this ML project is to predict customer churn. We are providing you a public dataset that has customer usage pattern and if the customer has churned or not. We expect you to develop an algorithm to predict the churn score based on usage pattern. The predictors provided are as follows:

account length:

total evening calls

total evening charge

international plan

total night minutes

total day minutes used

total night calls

voice mail plan

number of voice mail messages

number of customer service calls made

total night charge

day calls made

total international minutes used

total day charge

total international calls made

total evening minutes

total international charge

DataSet Information

The dataset describes mainly a how the customer churn is dependent on various factors e.g account length,total day calls,total evening calls,total day charge total night charge etc.

It has a total of 3333 observations.

Some of the factors like phoneno,areacode,state,international plan,voice mail plan may not be considered for predicting customer churn and hence shouldn’t be considered

The target class which is customer churn is having a value of either true or false(0 or 1).

The following is the distribution of the customer churn of the 3333 observations provided:

```
Churn=Flase    2850
Churn =True     483
```

Which is around **16.9% of customer churn**.The idea is the idenfity the factors listed above which is impacting customer churn so that appropriate business actions are taken based on these factors

Solution Statement

The objective of this ML project is to predict customer churn**.** Since it’s a labelled data we can use we can use supervised learning and also the problem statement is to predict customer churn hence it’s a classification problem.

Benchmark Models

Since it’s a classification problem,the following classification techniques can be used.

1.Logistic Regression

2.Random Forest.

3.Gaussian Naïve Bayes.

4.SVM

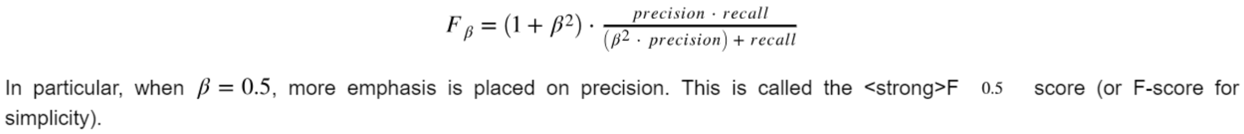
5.Ensemble Technique

The problem statement is about doing research on individuals those are likely to churn as operators/service providers would be interested on these individuals.Because of this assumption,the operators would be interested in how customer churn is predicted **accurately.**

Note there is imbalance in the data(churn v/s no churn) hence **accuracy** as a metric for evaluating a particular model performance won’t be appropriate.

Additionally,identifying someone who doesn’t churn as someone who churn would be detrimental since operators are trying to individuals who can churn.Therefore a model’s ability to precisely predict those who actually churn is more important than the model’s ability to recall these individuals.

We can use F-beta score as a metric that considers both precision and recall



$$F_{\beta} = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

In particular, when $\beta = 0.5$, more emphasis is placed on precision. This is called the <strong>F $_{0.5}$ score (or F-score for simplicity).

Looking at the distribution of classes(churn/not churn),its clear that most individuals doesn't churn.Hence its good to start with a statement that all customers would not churn. Making such a statement is **naïve** ,since no information is considered to substantiate this claim.It is always important to consider the naïve prediction for your data,to establish a benchmark for whether a model is performing well.

Evaluation Metrics

The following metrics can be considered for choosing the optimal classification model

1. F score

2.Training Time

3.Prediction Time

4.Cross Validation Results

As explained in the above section **"BenchMark Models"** its all about predicting whether the customer would churn or not one metric to think about is accuracy score (t determine how accurately the customer churn can be predicted),there is a imbalance in the data(churn v/s no churn) hence it wouldn't be good idea rather its always good to know the models ability to precisely predict the customer churn or recall of these customers the **F-score which is combination of precision or recall would be an important metric to consider**

The **Cross validation results** is all about how the model is performing on a train and test data with a train and test data split(there shouldn't be huge difference in the accuracy/f scores for both train and test data split)

The **training and prediction time** metrics would also be a deciding factor for model selection.

In case training and prediction time is given preference over accuracy we can compromise slightly on accuracy in decide on model selection based on the training and prediction time.

The metrics I would use to compare models would be **F score**.F score which is combination of both precision and recall would be a important metrics to consider.Accuracy can’t be considered as the metrics to compare models as there is imbalance in data(churn v/s no churn)

Project Design

I follow the following sequence of steps

1. Data Visualization: Visually representing data to find correlations between attributes and target variables.

This will also help finding the visible patterns in the data set.

2.Prepocessing the data:Scaling and Normalizing the data.As I will use the data available on the UCI ML repository and

not the one provided by the author on GitHub.I have to split the data into training,validation and testing data sets.

3.Feature Engineering:Finding relevant features,engineering new features etc.

4.Model Selection:Experiment with various models and algorithms to find the best one.

Ex Multi-layer perceptron which has not been used by the author.Others are of course regression techniques,SVM and gradient boosting

5. Model Tuning:Fine tune the selected algorithm to increase performance while making sure it doesn't overfit.

6.Testing:Test the model on the testing dataset.

References

- https://www.kaggle.com/becksddf/churn-in-telecoms-dataset/data
- https://www.crowdanalytix.com/contests/why-customer-churn/
- https://github.com/ctufts/Cheat_Sheets/wiki/Classification-Model-Pros-and-Cons
- https://ieeexplore.ieee.org/document/8258400/
- https://ieeexplore.ieee.org/document/5705218/?reload=true