2021

PREDICTING CUSTOMER CHURN RATE

TASK 1

AMAAN SHAIKH

Abstract

This project explores how customer churn can be predicted using machine learning models. While customer churn occurs in all types of industries, the model I created used sample data from a firm in the telecom industry and the objective was to predict whether the customer was going to continue using the same telephone provider or leave. It can be profitable for a company to find out which customers are predicted to churn due to the higher cost of acquisition of customers compared to retention. Loyal customers are also more likely to repurchase and refer the product. Models were trained using classification algorithms such as logistic regression, decision tree and an ensemble model but only the result of the model with the best metrics were included in the report.

1. **Problem Statement**

   Churn prediction is probably one of the most important applications of data science in the commercial sector. Churn rate, sometimes also called attrition rate, is the percentage of customers that stop utilizing a service within a time given period. It is often used to measure businesses which have a contractual customer base, especially subscriber-based service models. Using machine learning and data science to predict customer churn rate is tangible to comprehend and it plays a major factor in the overall profits earned by the business. If we manage to identify the customers who are going to churn and take the appropriate actions to avoid it in an automated fashion, then the profitability of the company is going to improve considerably. For this project, I will be focusing on the telecom industry by using a sample telco customer churn dataset to train a model, the dataset is available on my GitHub project repository (https://github.com/dpsfighters/Feynn-Labs-first-project).

2. **Market/Customer/Business Need Assessment**

   There are various industries and business models where companies can take advantage of predicting customer churn. These industries include retail, finance, telecom, travel, online retail, and any business with a subscription model.

Customer churn is a major problem and one of the most important concerns for large companies. Due to the direct effect on the revenues of the companies, especially in the telecom field, companies are seeking to develop means to predict potential customer to churn.

In relation to the Telco dataset, if we assume that each customer pays $60 and each month 14,500 customers churn, which means that we reduce our monthly revenue by $870,000 ($60 x 14,500 churned customers). If we manage to identify the customers who are going to churn and take the appropriate actions to avoid it in an automated fashion, then the profitability of the company is

going to improve considerably. Assuming that our CAC is 50$ and that the retention cost is five times lower ($10), if we reduced the churn rate by a 50%, we would save $373k (($60 - $10) * 7,250 retained customers) in just the first month.

3. **Target Specifications and Characterization**

   To help businesses with customer churn, a machine learning model can be created to predict which customers are likely to churn. Therefore, finding factors that increase customer churn is important to take necessary actions to reduce this churn. The main contribution of our work is to develop a churn prediction model which assists telecom operators to predict customers who are most likely subject to churn. By finding these customers, we can reach out to them with special offers or discounts to keep them engaged with the business. By doing this, the business will be making a greater profit overall instead of not taking any action at all.

4. **External Search (information sources/references)**

   To create a machine learning model, I used a sample dataset which can be found on my GitHub project repository (https://github.com/dpsfighters/Feynn-Labs-first-project). I will also be using a platform called BigML to create the machine learning models so there won't be any additional code on the GitHub repository.

5. **Bench marking alternate products**

There are other products and services that provide a similar service. There are many businesses, big and small, which provide data science and machine learning services. Nowadays, most companies have a small data science department or business analysts who provide these type of machine learning solutions. There are also consulting firms out there who provide these type of services to companies. There are also ways for someone to independently create their own models using softwares such as Python and R as they have existing libraries and packages such as TensorFlow, NumPy, KernLab etc.

6. **Applicable Regulations (government and environmental)**

The would be the regulations for a firm providing data science solutions

- Data protection and privacy regulations (Customers)
- Government Regulations
- Ownership/protection laws
- Antitrust/competition laws
- Data regulation

7. **Applicable Constraints (need for space, budget, expertise)**

There are various constraints that the business will have to consider

- Data Collection and storage (online servers, third party websites)
- Training and hiring data scientists
- Educating and training users to get familiar with the platform
- Marketing and reaching out to businesses to use the service

8. **Business Opportunity**

A website or business could be created where companies seeking assistance in predicting customer churn can approach us and we would assist them for adequate compensation. The company would then provide us data about its past and current customers which would then be used in machine learning model to predict which current customers are currently at risk of churning. This can be a viable business opportunity due to the fact that customer churn exists in many different industries and in companies big and small.

9. **Concept Generation and Development**

   This business opportunity requires creating machine learning models from scratch in order to cate to the needs of the client. However, for clients from similar industries, existing machine learning can be repurposed.

   A user interface would need to be created which can be in the form of a website where the client can request for a quote and then their needs can be catered to by a member of the business. However, a long-term goal would be to integrate the ability to make machine learning models on the website itself and the client themselves would enter the data and the target variable.

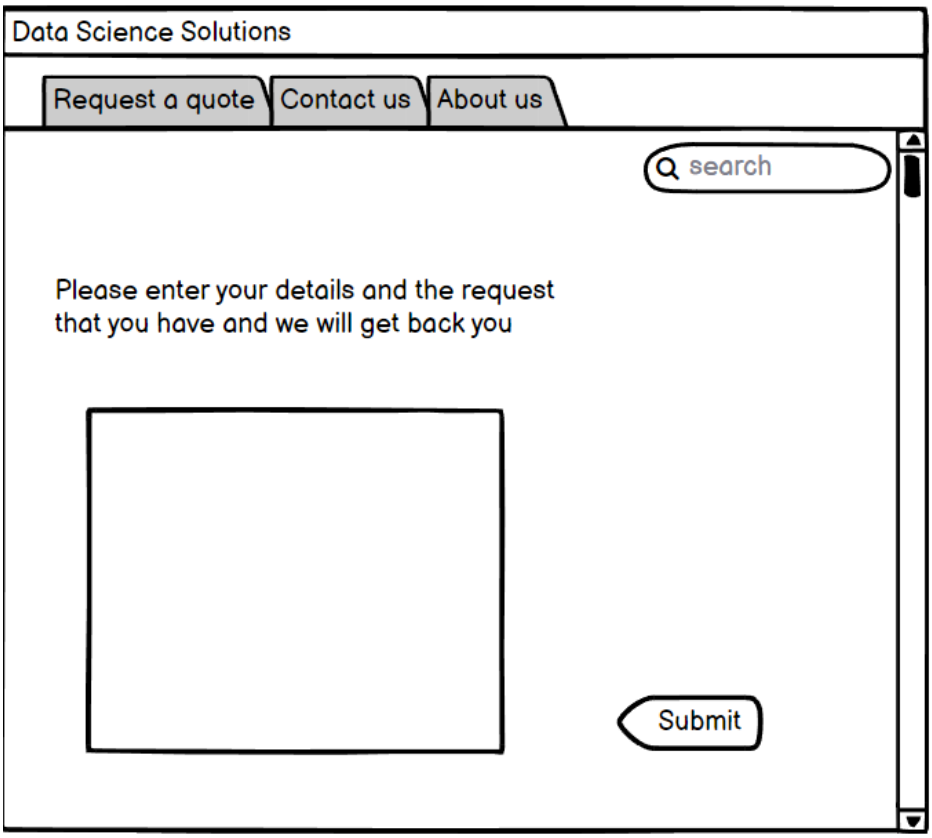
10. **Final prototype**

    •A client would go onto the website and enter the contact details to request for a quote

    •An analyst from the business would then respond to the client to gather additional information such as data about customers to build a machine learning model.

    •The client would then receive the results of the model's prediction telling which customers are predicted to churn so they can take the appropriate actions.

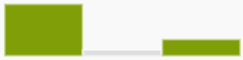
This is a prototype of what the website could look like. However, a long-term goal would be to make a platform where customers themselves can enter the data to create their models and this wouldn't be limited to just predicting customer churn, it can branch out to other machine learning problems.



11. Product Details and ML modelling

For the machine learning model, I will be using a platform called BigML which is an online platform where people can transform data into actionable models.

In relation to the telco churn dataset, we want to determine whether a customer will churn. Since there are a finite number of outcomes (Churn or No Churn), this is a classification problem. We have the previous customers' data and their outcomes. In other words, we have data about their past behavior, and we also know if they left the service or not. So, we will train a model to learn how the data relates to an outcome. Therefore, it is a Supervised problem. The data set has 3342 instances and there are 4 categorical and 16 numerical fields. However, I first cleaned the dataset and removed some anomalies which reduced the number of instances to 3276. The target variable is “churn” as we want to know whether a customer would churn or not. Here is an overview of the dataset:

Name	Type	Count	Missing	Errors	Histogram
State	ABC	3,276	0	0	
Account length	123	3,276	0	0	
Area code	123	3,276	0	0	
International plan	ABC	3,276	0	0	
Voice mail plan	ABC	3,276	0	0	
Number vmail messages	123	3,276	0	0	
Total day minutes	123	3,276	0	0	
Total day calls	123	3,276	0	0	
Total day charge	123	3,276	0	0	
Total eve minutes	123	3,276	0	0	
Total eve calls	123	3,276	0	0	
Total eve charge	123	3,276	0	0	
Total night minutes	123	3,276	0	0	
Total night calls	123	3,276	0	0	
Total night charge	123	3,276	0	0	
Total intl minutes	123	3,276	0	0	
Total intl calls	123	3,276	0	0	
Total intl charge	123	3,276	0	0	
Customer service calls	123	3,276	0	0	
Churn	ABC	3,276	0	0	

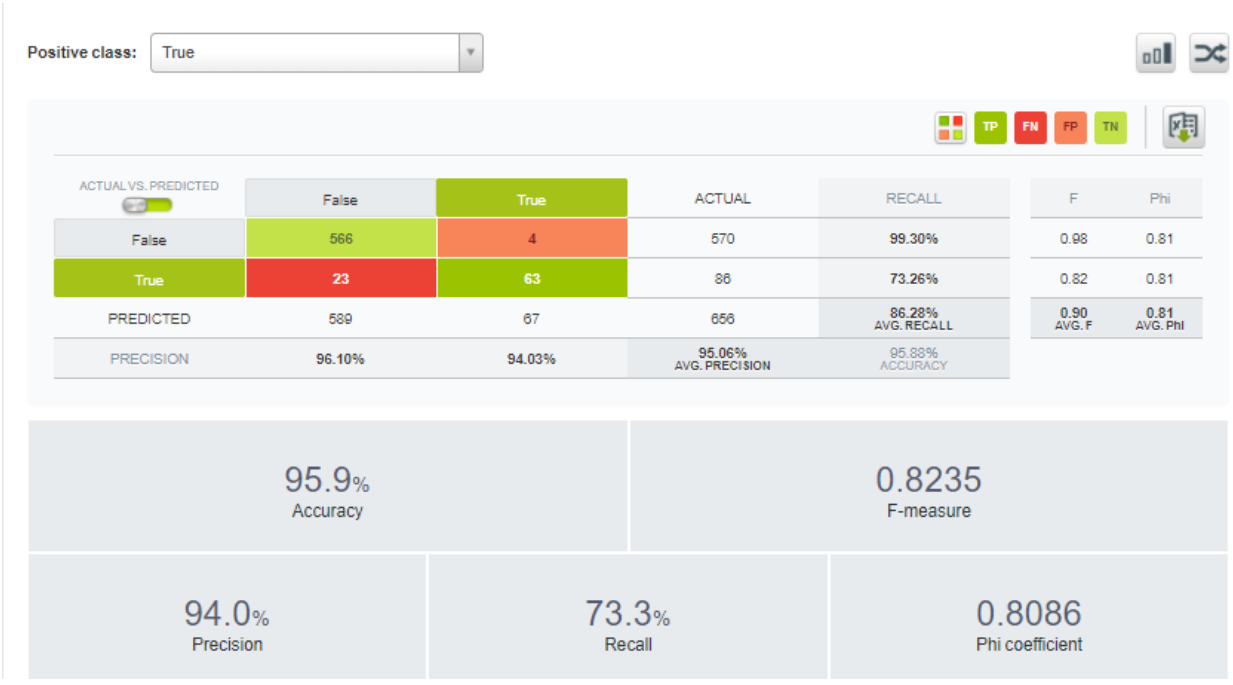
There are some variables which don't seem to be that important such as "State", "Account length" and "Area code" and I will not be considering them in my models as they won't be important in predicting the target variable. After splitting the data set into train and test, I trained different models using logistic regression, decision tree and an ensemble model but I displayed the results of the model which had the best metrics. That model was an ensemble of 25 random trees and here is the feature importance according to the model:

```
Data distribution:
    False: 86.73% (56810 instances)
    True: 13.27% (8690 instances)


Predicted distribution:
    False: 86.83% (56871 instances)
    True: 13.17% (8629 instances)


Field importance:
    1. Total day charge: 27.43%
    2. Customer service calls: 16.12%
    3. Total eve charge: 12.24%
    4. International plan: 9.48%
    5. Number vmail messages: 6.39%
    6. Total day minutes: 4.37%
    7. Total intl charge: 4.26%
    8. Total day calls: 4.01%
    9. Total night calls: 3.67%
    10. Total night minutes: 3.53%
    11. Total intl calls: 3.25%
    12. Total eve calls: 3.23%
    13. Total eve minutes: 1.19%
    14. Total intl minutes: 0.43%
    15. Total night charge: 0.39%
```

It is reasonable that the variables referring to (1) the use of the service (such as the charges, whether the customer is on an international plan and, to a lower extent, the number of calls) and the one representing (2) the amount of problems suffered by customers (customer service calls)are the ones with the highest predictive power. A human expert surely would agree with most of them. Here is the confusion matrix of the model:



- True Positive = Correct identification of a customer who is going to churn.
- True Negative = Correct identification of a customer who is NOT going to churn.
- False Positive = Classifying a loyal customer as a churner.
- False Negative = Classifying a churner as a loyal customer

It can be a bit difficult to interpret the metrics at first:

Accuracy = Correct predictions over all data points = (TP + TN) / All data points =95.9% —> Apparently, this result seems extremely good! We are able to predict correctly 95.9% of all the data points.

Recall = Share of Churners which are properly predicted = TP / (TP + FN) = 73.3% —>This results is not as good as the accuracy… —> We see that this Ensemble tends to predict Churners as Loyal customer —> 23 False Negatives.

Precision = Share of correctly predicted Churners over all customers predicted as Churners = TP / (TP + FP) = 94% —> This metric looks good! The model does not predict many Loyal customers as Churners. We would like to add Specificity = Share of Loyal customers which are properly predicted = TN / (TN + FP) = 94% —> Good performance, like Precision

Out of all the metrics, I would be paying more attention to the metric recall. Accuracy isn't an important metric as the target variable is imbalanced (high a number of no compared to yes for churn rates) as a high accuracy can be obtained by classifying all observations as the majority class and therefore accuracy is more useful when the target variable is more equal.

The F1 score is the harmonic mean of precision and recall. It may be better than Accuracy because it is not biased by a large number of True Negatives, such as in this case. However, it is difficult to interpret. e.g., What does an F1 score of 0.8 mean? It only allows us to rank the different models. It assumes that False Positives and False Negatives have the same value, which is usually not the case, as in our example.

Precision measures how accurate the model is in terms of finding the number of actual positives from predicted positive instances and is useful when the cost of false positives is high. Recall measures the ratio of true positives to total (actual) positives in the data and it is the most important to us. This metric places more importance on the false negatives which is important to

the case as it is much more the company would rather have some customers who are predicted to churn but stay (false positive) rather than customers who are predicted to stay but churn (false negative) as these are the customers who will be lost.

Based on the feature importance of the model, once we identify a customer who is going to churn, the following actions can be taken be taken by the client.

- Offer a temporary discount on the price of day minutes.
- Send SMS during key holidays (e.g., Christmas, Thanksgiving, Easter…) reminding customers to call their loved ones.
- Offer a discount on the International Plan.
- Offer a free trial of the International Plan during next summer.
- Analyze the customer service calls, understand which are the main pain points and solve them.

12. Conclusion

Predicting customer churn is an important decision that many businesses should take due to the high cost of acquisition of new customers. Acquiring a new customer can be up to five times more expensive than retaining an existing customer and increasing customer retention by 5% can increase profits drastically. This is why using machine learning to predict customer churn can be such a huge asset to firms in such industries due to the number of losses that can be recouped.

References

https://www.globallegalinsights.com/practice-areas/ai-machine-learning-and-big-data-laws-and-regulations/india

https://towardsdatascience.com/predict-customer-churn-the-right-way-using-pycaret-8ba6541608ac

https://www.outboundengine.com/blog/customer-retention-marketing-vs-customer-acquisition-marketing/

https://www.analyticsvidhya.com/blog/2021/08/churn-prediction-commercial-use-of-data-science/

https://www.statista.com/statistics/816735/customer-churn-rate-by-industry-us/

https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6