1. INTRODUCTION

1.1 Overview

Customer churn is a major problem and one of the most important concerns for large companies. Telecommunication industry always suffers from very high churn rates when one industry offers a better plan than the previous there is a high possibility of the customer churning from the present due to a better plan in such a scenario it is very difficult to avoid losses.[1]

This project takes part of ‘Build-A-Thon’, a collaboration between IBM, Smart Internz, and IEEE Sup’Com. During this project, I will be doing one of the projects that are proposed by Smart Internz through their platform. I chose to work on the Telecom Customers Churn project because of its challenging nature. Eventually, I will be building a machine learning model that predicts if a given customer of a telecommunications company will maintain its subscription to the company’s services or it will withdraw it based on its coordinates that are given through and interface.

1.2 Purpose

Due to the direct effect on the revenues of the companies, companies are seeking to develop means to predict potential customers to churn. Therefore, finding factors that increase customer churn is important to take necessary actions to reduce it.

Throughout this particular project, Telecommunications operators will have an insight on who among their customers will be dropping its subscription to their services given its credentials and coordinates such as Region, Income, Ownership of Credit Card, …

2. LITERATURE SURVEY

2.1 Existing problem

Here are the existing approaches to solve this particular problem currently :

a. Listen to what your customers are saying

Staying in sync with what your customers expect from you is the foundation to an effective customer retention strategy. And to stay in sync, you need to listen to customers across multiple touchpoints. You need to be available for them, at their convenience so that you can record their feedback, specific queries and complaints at any particular time.

b. Don’t just be open to feedback. Action it!

Most churn doesn’t happen overnight. It’s mostly the end outcome of an accumulation of events where the customer feels unhappy with your services. Which is why you need to attend to it at the very beginning and not let the dissatisfaction grow. This translates to acting on customer feedback on time, appeasing disgruntled customers and closing the loop effectively. [2]

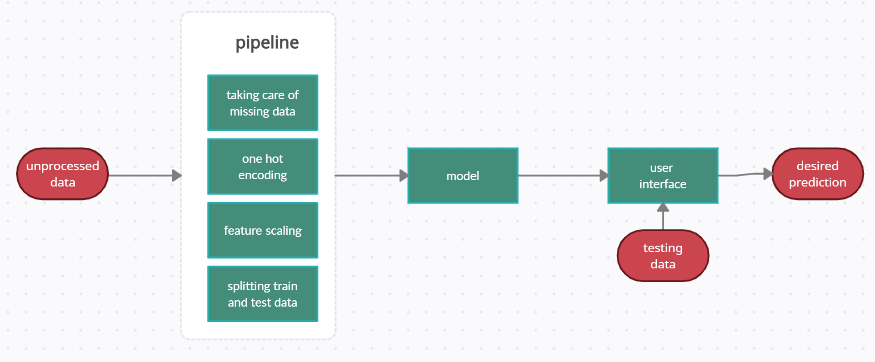
2.2 Proposed solution

I proposed a method which will let Telecommunications operators have an insight on who among their customers will be dropping its subscription to their services given its credentials and coordinates such as Country, Income, Ownership of Credit Card, etc. This is done through a machine learning model which can be used through a user-friendly interface.

3. THEORITICAL ANALYSIS

3.1 Block diagram

Diagrammatic overview of the project.



3.2 Hardware / Software designing

For this project, I used a handful of tools to get my work done. Here are the tools that I used:

Kaggle: it is a platform in which there is the required dataset as well as notebooks of people who have already worked on this project which I did inspect to take a general idea on how to approach this particular problem even though it wasn’t very beneficial since I had to go by well specified steps in the project description (I wrote the entire code by myself to match the project’s steps description)

IBM Cloud: Using IBM Cloud’s public platform, I created three vital services for the projects who are respectively; IBM Cloud Object Storage Service, IBM Watson Studio Service, and an IBM machine learning service. The first two tools are crucial for the functionality of the last, which we will talk about further in detail;

IBM Machine Learning Service: It is a service that is provided by IBM as previously mentioned. In this service, you can create a project and allocate certain resources to launch the project. In the given project instance, I uploaded the required dataset to put it into use in a notebook that is created in this project as well.

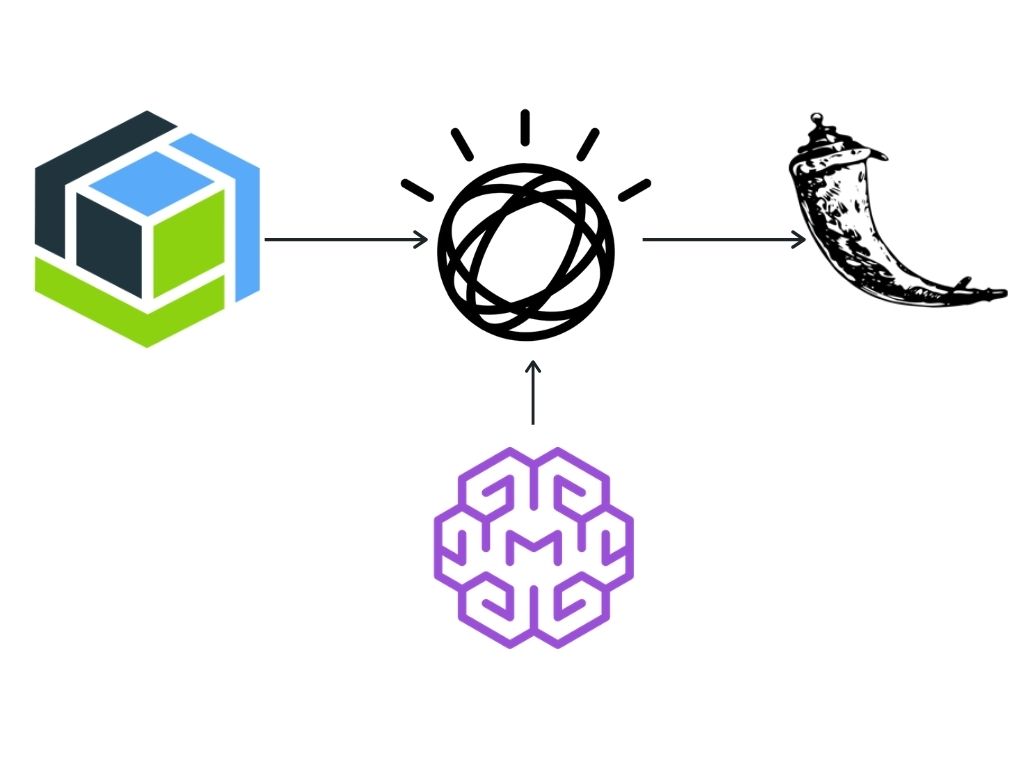
VS Code: it is a text editor that I used to build the flask application as required within the project’s steps’ description

Flask: it is a web development framework that is light and machine-learning oriented.

4. EXPERIMENTAL INVESTIGATIONS

Throughout the project, I tried at first to get insights about the customer churn problem in general to get a general view on the matter and to specify the desired output ; most of the data provided to Telecom companies are related to its actions like its level of involvement or the amount of money spent on credit in a certain period of time, and very few parameters are actually insightful on this customer really is such as name (which doesn’t provide much insight already), age , region or country, ... The point of this is to determine as much insightful data on the customer to ease the work on the pipeline and the model in general since the right amount of insightful data you give to the model, the less it is likely to overfit or underfit. I also read a lot of state-of-art literature on Machine Learning models to specify the adequate one to work with. Two main models were the best candidates for this binary classification problem which are ANNs and Xgboost. I chose to work with Xgboost because it’s an ensemble method and according to the ‘Hands on Machine Learning using SickitLearn and TensorFlow’[3], the xgboost model wins most of Kaggle’s competitions due to its easy understanding and the efficiency of its hyperparameters tuning using preconfigured tools from different libraries. So xgboost it is! After this, I started thinking about the pipeline (which was already predescribed in the project’s workspace in SmartInternz’s website, so I just had to assemble it.

5. FLOWCHART



6. RESULT

The solution that I proposed gave around 86% accuracy, it is not very good but given the short amount of time given to bring this project into reality, and this is what I could’ve come up with. In the end of the model’s script, I putted a confusion matrix which gave a 73.75% precision and 51.35% recall which I admit isn’t the best possible output, but at least the precision is significantly higher than recall (more than 20% difference) so at least I’m on the right track. Also, the solution is immune against overfitting because even though in hyperparameter tuning, I putted many possible hyperparamters, the length of the dataset and the number of columns are sufficient to avoid overfitting so that’s pretty good.

Now for the interface; it is simple to use and has good basic estatics to not overwhelm by complicated design.

Here are some screenshots for interface before entering data instance for testing and after it:

7. ADVANTAGES & DISADVANTAGES

This proposed solution has various aspects that range between positive and negative:

Advantages: it substitutes previously proposed solutions such as listening to customer’s feedback individually which is resource-consuming (time, budget) and it is easily exposed to bias because it depends on the customer’s capability to deliver an honest and an optimal feedback which is not the case mostly because humans have different expression methods and it depends on things such as mood and environment. So this proposed solution is much much better in terms of liability. It also comes with a nice user interface to ease the work of regular user.

Disadvantages: the given dataset is very limited in terms of ‘width’ and ‘length’; there simply isn’t enough columns because there obviously other factors that are critical to know if the given customer will maintain subscription or withdraw it such as date of subscription, the amount of money spent on credit by that customer, … the given solution also is very limited in terms of user usability; you can only enter new data instances one by one via the interface. The proposed solution also uses off-line training so it really doesn’t change the model as time goes by. This, of course, goes back to the fact that if I wanted to implement a model that is dynamic such as LSTM, it would take time to the conception but the temporal constraint simply won’t let me finish it.

8. APPLICATIONS

This of solution, obviously, will be applied for telecommunications companies that want to work on their customer service by detecting customers that will withdraw subscription. This will help them precautionary measures either on the individual level of each customer, either on collective level; they can detect the pattern that goes by among customers that will withdraw subscription and know why they do so, or at least form the persona of the customer that will withdraw the subscription anyway no matter how ‘spoiled’ it get so the company don’t waste resources on that particular customer.

9. CONCLUSION

In conclusion, this project is a customer churn prediction model that will help telecommunications companies optimize their strategies (technical strategies, marketing strategies, …). The service of this project is provided through a user interface.

10. FUTURE SCOPE

Many enhancements can be made in the future;

Technical aspect: the user interface can become more useful by adding the option of adding data for testing by the batch through an excel sheet or a csv file (and the output will be the same of course). Also, the model can be improved by using LSTM to make it dynamic.

Functional aspect: the model can be modified to become a multilabel classification problem; there are three classes of customers; the first are customers that will maintain subscription so we don’t have to worry about them. The second are the customers that will withdraw subscription but if we will give them certain privileges, such as extra credit or a small gift upon spending a given amount of money, that customer will maintain subscription so it’s in the best interest of the company to keep spoiling that customer. The third class are the customers that will withdraw subscription no matter the company will do to them, so it’s wasteful to allocate resources for this range of current customers. To do this, a different dataset is required.

11. BIBILOGRAPHY AND WEBIOGRAPHY

[1] : <https://smartinternz.com/Student/badge_workspace/8834>

[2] : <https://cloudcherry.com/blog/improve-customer-retention-in-telecom/>

[3] : <http://index-of.es/Varios-2/Hands%20on%20Machine%20Learning%20with%20Scikit%20Learn%20and%20Tensorflow.pdf>

APPENDIX

A. Source Code