

**Bharat Institute of Technology**

Department of Computer Science and Engineering

A project report on CHURN PREDICTION

|  |  |  |
| --- | --- | --- |
| **Submitted to:** | **Submitted By:** |  |
| Aditya Bhusan  (1612810008) |  |
| Mr. Kshitiz Saxena |  |
| Aditya Ranjan  (1612810009)  Aditya Ranjan  (1612810010) |  |
| Head Of Department, |  |
|  |  |
| Department of Computer Science and Engineering, |  |  |
| Bharat Institute Of Technology,Meerut. |  |  |

P a g e | **2**

**Certificate of the Project**

This project is for the department of Computer Science & Engineering, under the

faculty of Engineering & Technology at Bharat Institute Of Technology, in partial

fulfillment of the requirements for the degree of Bachelor of Science in Engineering in Computer Science & Engineering in 2020 academic year.

**Mentor :**

Mr. Anubhav Sharma

Department of Computer Science and Engineering,

Bharat Institute Of Technology,Meerut.

**Author :**

|  |  |
| --- | --- |
| **Student Name** | **Student ID** |
|  |  |
| Aditya Bhusan | 1612810008 |
|  |  |
| **Aditya Ranjan**  **Aditya Yadav** | **1612810009**  **1612810010** |
|  |  |

Department Of Computer Science and Engineering

Bharat Institute Of Technology,Meerut

P a g e | **3**

**Declaration**

I the undersigned solemnly declare that the project report on Churn

Prediction is based on my own work carried out during the course of our

study.

|  |
| --- |
| Aditya Bhusan  (1612810008) |
|
| Aditya Ranjan  (1612810009)  Aditya Ranjan  (1612810010) |
|

**Acknowledgement**

It is my great pleasure to express my profound sense of gratitude to my Supervisor Mr. Kshitiz Saxena, Head of Department of Computer Science and Engineering,

Bharat Institute Of Technology for his constructive academic advice and guidance, constant encouragement and valuable suggestions, and all other supports throughout this

project work and preparing this project report successfully.I am really benefited from his excellent supervision.

I would like to thanks to all of our friends and those who helped, inspired and gave us mental support at different stages in different moment in our project.

I would also like to thank my mentor Mrs. Sakshi Jain for mentoring and guiding us throughout the project.

**Acknowledgement**

It is my great pleasure to express my profound sense of gratitude to my Supervisor Mr. Kshitiz Saxena, Head of Department of Computer Science and Engineering,

Bharat Institute Of Technology for his constructive academic advice and guidance, constant encouragement and valuable suggestions, and all other supports throughout this

project work and preparing this project report successfully.I am really benefited from his excellent supervision.

I would like to thanks to all of our friends and those who helped, inspired and gave us mental support at different stages in different moment in our project.

I would also like to thank my mentor Mrs. Sakshi Jain for mentoring and guiding us throughout the project.

P a g e | **4**

**Abstract**

The purpose of the project entitled as “Churn Prediction” is to computerize "Churn Rate" is a business term describing the rate at which customers leave or cease paying for a product or service.Consequently, there's growing interest among companies to develop better churn-detection techniques, leading many to look to data mining and machine learning for new and creative approaches. This is a post about modeling customer churn using Python.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  | P a g e | **5** | |
|  |  |  |  | |  |  |  |
|  |  |  | **TABLE OF CONTENTS** | |  |  |  |
|  |  | | | | |  |  |
|  | **Chapter Name** | | | | | **Page No.** |  |
|  | | | | | |  |  |
| **Chapter One : INTRODUCTION** | | | | | | **5** |  |
| 1.1 Introduction | | | |  |  |  |  |
| 1.2 Objectives | | | |  |  |  |  |
|  | | | | | |  |  |
| **Chapter Two : LITERATURE REVIEW** | | | | | | **6-8** |  |
| 2.1 Technology Used | | | | | |  |  |
| \*Jupyter | |  |  |  |  |  |  |
|  | | | | | |  |  |
| **Chapter Three : METHODOLOGY** | | | | | | **9-11** |  |
| 3.1 | What is Churn Prediction | | | | |  |  |
| 3.2 features of Churn Prediction | | | | | |  |  |
| 3.3 | Software Requirements | | | | |  |  |
|  |  | | | | |  |  |
|  | | | | | |  |  |
| **Chapter Four : SCREENSHOT WITH PROJECT DETAIL** | | | | | | **12-15** |  |
| 4.1 | Main Interface | | |  |  |  |  |
| 4.2 | User Sign Up And Login Panel Interface | | | | |  |  |
| 4.3 | Admin Panel Interface | | | | |  |  |
|  | | | |  |  |  |  |
| **Chapter Five : ANALYSIS** | | | |  |  | **16-17** |  |
| 5.1 | The Convenience | | | Aspect | |  |  |
| 5.2 | The Merchandise | | | Aspect | |  |  |
| 5.3 | The Cost Aspect | | |  |  |  |  |
| 5.4 | Economic | Feasibility | | | |  |  |
| 5.5 | Technical | Feasibility | | | |  |  |
| 5.6 | Operational Feasibility | | | | |  |  |
|  | | | | | |  |  |
| **Chapter Six : Future Plan And Conclusion** | | | | | | **18** |  |
| 6.1 | Future Plan | | |  |  |  |  |
| 6.2 | Conclusion | | |  |  |  |  |
|  | | | | | |  |  |
| **Chapter Seven : BIBLIOGRAPHY** | | | | | | **19** |  |
| 7.1 | Book Reference | | |  |  |  |  |
| 7.2 | Website Reference | | | | |  |  |
|  |  |  |  |  |  |  |  |

P a g e | **6**

**CHAPTER 1**

**INTRODUCTION**

**1.1 Introduction :**

Churn is defined slightly differently by each organization or product. Generally, the customers who stop using a product or service for a given period of time are referred to as churners. As a result, churn is one of the most important elements in the Key Performance Indicator (KPI) of a product or service. A full customer lifecycle analysis requires taking a look at retention rates in order to better understand the health of the business or product.

In the gaming industry, churn comes in different flavors and at different speeds. For instance, in games where players must be engaged on a day-to-day basis, a player who doesn't login within 24 hours may be considered a churner. On the other hand, in games where players aren’t necessarily playing the game every day, the time frame that makes up a churn is much longer. It is important for a predictive pipeline to be robust enough to handle such variances.

From a machine learning perspective, churn can be formulated as a binary classification problem. Although there are other approaches to churn prediction (for example, survival analysis), the most common solution is to label “churners” over a specific period of time as one class and users who stay engaged with the product as the complementary class.

**1.2 Objectives :**

The project objectives are :

\*To predict the customers most likely to churn

P a g e | **7**

**CHAPTER 2**

**LITERATURE REVIEW**

**2.1 Technology Used**

Data set

There are many types of data in SyriaTel used to build the churn model. These types are classified as follow:

1. 1.

*Customer data* It contains all data related to customer’s services and contract information. In addition to all offers, packages, and services subscribed to by the customer. Furthermore, it also contains information generated from CRM system like (all customer GSMs, Type of subscription, birthday, gender, the location of living and more ...).

1. 2.

*Towers and complaints database* The information of action location is represented as digits. Mapping these digits with towers’ database provides the location of this transaction, giving the longitude and latitude, sub-area, area, city, and state.

Complaints’ database provides all complaints submitted and statistics inquiries related to coverage, problems in offers and packages, and any problem related to the telecom business.

1. 3.

*Network logs data* Contains the internal sessions related to internet, calls, and SMS for each transaction in Telecom operator, like the time needed to open a session for the internet and call ending status. It could indicate if the session dropped due to an error in the internal network.

1. 4.

*Call details records “CDRs”* Contain all charging information about calls, SMS, MMS, and internet transaction made by customers. This data source is generated as text files.

P a g e | **8**

1. 5.

Mobile IMEI information It contains the brand, model, type of the mobile phone and if it’s dual or mono SIM device.

This data has a large size and there is a lot of detailed information about it. We spent a lot of time to understand it and to know its sources and storing format. In addition to these records, the data must be linked to the detailed data stored in relational databases that contain detailed information about the customer. The nine months of data sets contained about ten million customers. The total number of columns is about ten thousand columns.

## Data exploration and challenges with SyriaTel dataset

Spark engine is used to explore the structure of this dataset, it was necessary to make the exploration phase and make the necessary pre-preparation so that the dataset becomes suitable for classification algorithms. After exploring the data, we found that about 50% of all numeric variables contain one or two discrete values, and nearly 80% of all the categorical variables have Less than 10 categories, 15% of the numerical variables and 33% of the categorical variables have only one value. Most of some variables’ values are around zero. We found that 77% of the numerical variables have more than 97% of their values filled with 0 or null value. These results indicate that a large number of variables can be removed because these variables are fixed or close to a constant. This dataset encounters many challenges as follow.

### **Data volume**

Since we don’t know the features that could be useful to predict the churn, we had to work on all the data that reflect the customer behavior in general. We used data sets related to calls, SMS, MMS, and the internet with all related information like complaints, network data, IMEI, charging, and other. The data contained transactions for all customers during nine months before the prediction baseline. The size of this data was more than 70 Terabyte, and we couldn’t perform the needed feature engineering phase using traditional databases.

### **Data variety**

The data used in this research is collected from multiple systems and databases. Each source generates the data in a different type of files as structured, semi-structured (XML-JSON) or unstructured (CSV-Text). Dealing with these kinds of data types is very hard without big data platform since we can work on all the previous data types without making any modification or transformation. By using the big data platform, we no longer have any problem with the size of these data or the format in which the data are represented.

### **Unbalanced dataset**

The generated dataset was unbalanced since it is a special case of the classification problem where the distribution of a class is not usually homogeneous with other classes. The dominant class is called the basic class, and the other is called the secondary class. The data set is unbalanced if one of its categories is 10% or less compared to the other one [[18](https://link.springer.com/article/10.1186/s40537-019-0191-6#CR18)].

Although machine learning algorithms are usually designed to improve accuracy by reducing error, not all of them take into account the class balance, and that may give bad results [[18](https://link.springer.com/article/10.1186/s40537-019-0191-6#CR18)]. In general, classes are considered to be balanced in order to be given the same importance in training.

We found that SyriaTel dataset was unbalanced since the percentage of the secondary class that represents churn customers is about 5% of the whole dataset.

### **Extensive features**

The collected data was full of columns, since there is a column for each service, product, and offer related to calls, SMS, MMS, and internet, in addition to columns related to personnel and demographic information. If we need to use all these data sources the number of columns for each customer before the data being processed will exceed ten thousand columns.

P a g e | **10**

**CHAPTER 3**

**METHODOLOGY**

**3.1 What is Churn Prediction?**

The volume of data has been growing at a fast rate over the

last two decades due to advancements in technology. Many

new methods and techniques have been introduced to process

the data and gather valuable information which is hidden in

the raw data. Data mining is defined as the process of

extracting valuable information from the data. Many data

mining methods have been successfully applied in various

domains.

Customers are the most important assets in any industry since

they are considered as the main profit source. Nowadays,

companies have become observant that they should put much

effort not only to convince the customers, but also to retain

their existing customers. Churners are persons who move to

other company for various reasons. To reduce customer

churn, the company should be able to predict the behaviour of

customer correctly and establish connections between

customer attrition and keep factors under their control. Churn

prediction is a binary classification task, which differentiates

churners from non-churners.

Customer churn is defined as the movement of people from

one bank to other bank. The main reasons for churn are

dissatisfaction with the customer service, high costs,

unattractive plans, bad support. It is an expensive problem in

many sectors since acquiring new customer costs five to six

times more than retaining existing ones

**3.2 Features of Churn Prediction :**

The data was processed to convert it from its raw status into features to be used in machine learning algorithms. This process took the longest time due to the huge numbers of columns. The first idea was to aggregate values of columns per month (average, count, sum, max, min ...) for each numerical column per customer, and the count of distinct values for categorical columns.

Another type of features was calculated based on the social activities of the customers through SMS and calls. Spark engine is used for both statistical and social features, the library used for SNA features is the Graph Frame.

* *Statistics features* These features are generated from all types of CDRs, such as the average of calls made by the customer per month, the average of upload/download internet access, the number of subscribed packages, the percentage of Radio Access Type per site in month, the ratio of calls count on SMS count and many features generated from aggregating data of the CDRs.

Since we have data related to all customers’ actions in the network, we aggregated the data related to Calls, SMS, MMS, and internet usage for each customer per day, week, and month for each action during the nine months. Therefore, the number of generated features increased more than three times the number of the columns. In addition, we entered the features related to complaints submitted from the customers from all systems. Some features were related to the number of complaints, the percentage of coverage complaints to the whole complaints submitted, the average duration between each two complaints sequentially, the duration in “Hours” to close the complaint, the closure result, and other features.

The features related to IMEI data such as the type of device, the brand, dual or mono device, and how many devices the customer changed were extracted.

We did many rounds of brainstorming with seniors in the marketing section to decide what features to create in addition to those mentioned in some researches. We created many features like percentage of incoming/out-coming calls, SMS, MMS to the competitors and landlines, binary features to show if customers were subscribing some services or not, rate of internet usage between 2G, 3G and 4G, number of devices used each month, number of days being out of coverage, percentage of friends related to competitor, and hundred of other features.

visualize some of the basic categorical and numerical features to give more insight on the deference between churn and non-churn classes.

**3.3 Software Requirement (any one) :**

* Jupyter

P a g e | **11**

**FUTURE PLAN AND CONCLUSION**

CONCLUSION

This paper provides a detailed study on the methods used for

the process of customer churn prediction. Each of the above

churn prediction models has low accuracy and prediction.

Hence a good prediction model is required in order to avoid

the customer churn problem. This can be achieved by

combining SVM with boosting algorithms for higher accuracy

and performance which can be considered as a future work for

Churn prediction. Good prediction models have to be

constantly developed and a combination of the proposed

methods has to be used.

P a g e | **20**

**CHAPTER 7**

**BIBLIOGRAPHY**

**7.1 Website Reference :**

* <https://archive.ics.uci.edu/ml/index.php>
* <https://archive.ics.uci.edu/ml/datasets.php>