Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

SUMMARY PROJECT PLAN

SUPERVISOR – MERIEL HUGGARD STUDENT NAME-SHRUTI KATHURIA TCD ID-21355061

Customer Churn Prediction using Synthetic Data

INTRODUCTION

Customer Churn Prediction is a process of predicting whether a customer is likely to stop doing business with a company. This is important for businesses as it allows them to proactively address the issue and take steps to retain the customer. The goal of a customer churn prediction project is to create a model that can accurately predict which customers are at risk of churning so that the company can take action to retain them. The project typically starts with the collection and pre-processing of data followed by Data Selection and then we apply some basic pre-processing techniques to data. We need to create synthetic data from it so we can increase the size of data which will be beneficial in terms of correctly model training. This data may include information about the customer's demographics, transaction history, and engagement with the company. Once the data is cleaned and prepared, various machine-learning algorithms can be applied to train a model to predict customer churn. The model is then evaluated using metrics such as accuracy, precision, and recall to ensure that it is performing well. In order to improve the model's performance, various techniques such as feature engineering, hyperparameter tuning, and ensemble methods can be applied. Overall, a customer churn prediction project can be an effective way for companies to identify customers at risk of leaving, so they can take steps to retain them. It's important for the company to have a good understanding of the data, feature engineering, and various machine learning algorithms to build an accurate model and make sure to evaluate it using the appropriate metrics.

Goal and Objective of the project:

The problem statement of customer churn prediction is to predict which customers are likely to cancel their subscriptions or stop using a company's products or services in the near future. This can help a company take proactive measures to retain these customers and prevent the loss of revenue. The goal of this project is to predict which customers are likely to churn (cancel their subscription or stop using a service) in the near future. This information will be used to target retention efforts and improve customer retention rates. The model will be trained on historical customer data, including demographic information and past interactions with the company. The performance of the model will be evaluated using metrics such as accuracy and AUC-ROC score and will be compared to a baseline model.

IMPLEMENTATION PHASE I

Data Selection

The first and essential step of machine learning model training is to collect past data for training the machine learning model on that data. So, I have selected the (BankChurners.csv) dataset from Kaggle for training my machine learning model.
The link to the dataset is here: https://www.kaggle.com/code/shiviyadav/customer-churn-prediction-eda-ml-techniques/data?select=BankChurners.csv

About the Dataset

Customer attrition is a challenge for business management with a portfolio of consumer credit cards. In order to anticipate which clients are most likely to stop buying from them, they want to analyses the data to determine the cause of this. Here are the dataset variables.

i. **CLIENTNUM –**
Client number, Unique identifier for the customer holding the account

ii. **Attrition_Flag**
Internal event (customer activity) variable - if the account is closed then 1 else 0

iii. **Customer_Age**
Demographic variable - Customer's Age in Years

iv. **Genders**
Demographic variable - M=Male, F=Female

v. **Dependent_count**
Demographic variable - Number of dependent

vi. **Education_Level**
Demographic variable - Educational Qualification of the account holder (example: high school, college graduate, etc.)

vii. **Marital_Status**
Demographic variable - Married, Single, Divorced.

viii. **Income_Category**
Demographic variable - Annual Income Category of the account holder (< $40K, $40K - 60K, $60K - $80K, $80K-$120K).

ix. **Card_Category**
Product Variable - Type of Card (Blue, Silver, Gold)

x. **Months_on_book**
Period of relationship with the bank

Data Reading

The very basic step in model training is data reading you have to read data and explore the data to know what actually the data is. Data reading in Python refers to the process of importing data from a file or external source into a Python program for further manipulation or analysis. This can be done using built-in libraries such as the "CSV" library for reading CSV files or the "pandas" library for reading a variety of file types. The data can then be stored in a variable, such as a list or a data frame, for further processing.

Data Pre-processing

I used pre-processing on the data after reading it. Cleaning, converting and getting ready the data for analysis or modelling is known as data pre-processing. Given that raw data is frequently inaccurate, inconsistent, or irrelevant and may contain outliers or errors, it is a crucial phase in the data science process. Data pre-processing is needed to make the data more usable and to ensure that the results obtained from any analysis or modelling are accurate and reliable. Steps that are typically included in data pre-processing in Python are included.

i)Data cleaning
ii. Data transformation
iii. Data integration
iv. Data reduction
v. Data splitting
vi. Data augmentation

Generating Synthetic Data

Information that has been intentionally annotated is known as synthetic data. It is produced via simulations or computer algorithms. When genuine data is either unavailable or needs to be kept confidential due to personally identifiable information (PII) or compliance problems, synthetic data production is typically used.

Why Generating Synthetic data

The purpose of generating synthetic data in Python is to create a set of artificial data that can be used for a variety of purposes such as testing machine learning models, evaluating the performance of algorithms, or creating a sample dataset for exploration and visualization. It can also be useful when real data is not available or cannot be used for legal, ethical, or other reasons. Synthetic data can also be used to artificially inflate the size of a dataset to better train machine learning models.

IMPLEMENTATION PHASE II

After synthetic data, we will apply pre-processing on synthetic data like finding missing values and filling them, etc. After pre-processing we will do exploratory data analysis on data which is simply called EDA. During EDA we will find how our data act like relationships and the effect of columns on each other. Once everything is done, we train our model and get the best accuracy and get the best prediction of churning customers which could benefit our company. After EDA we need to separate features into dependent and independent variables. The dependent variable is a variable that will be predicted and independent variables are variables that predicts the dependent variable.
After this, we need to split our data into training and testing parts the training data is the data that is used to train our model, and test data is used to test the model. After this, we need to apply some algorithms to training data and find their accuracy one by one and compare them using data visualization. After model training, we need to apply K fold validation for checking our model accuracy with a decent method and checking their working and testing accuracy.

After training the model, we have to deploy it using the python web framework FLASK.
In the deployment process first of all I will save the model and then set up a flask environment. After Setting up the flask environment we need to do some designing of the page that contains some input fields when the user fills in the data our model gives the predicted results.

Update about the supervisor meetings

Had regular meeting with my supervisor regarding the project on Tuesdays. We discussed a number of things related to the project. Initial meetings were related to what topic to choose. Later one's I decided to work on anything related to ML and synthetic data. Moved on to looking for data sets to apply the same. Had discussions about using time scaling data or not. Finally decided upon a dataset related to customer churning. Everything is mentioned in the Gantt chart including every process of the project.
The Link to the Gantt Chart is attached: Here

Gantt chart: