Electronics & ICT Academy
Indian Institute of Technology, Guwahati

Post Graduate Certification in

Data Science

Customer Engagement Enhancement in BFSI Industry

Industry Grade Project – II: PG Certification in Data Science by E&ICT Academy IITG

Table of Contents

Aim of the Project

> The aim of the project is to build a Machine Learning Model to identify the most unengaged users in FinTech industry to help them implement business actions to convert them to engaged users.

Background

ParTech is an emerging subscription-based financial technology start-up that offers a range of services to its customers, including bank accounts, debit cards, currency exchange, stock trading, cryptocurrency exchange, and peer-to-peer payments. As with many subscription-based businesses (especially in the FinTech domain), one of the main areas of concern for the company has been customer engagement and increasing churn rate. They need assistance from a team of Data Scientists in quantifying success and driving their business teams towards an effective customer retention strategy.

As Data Scientists, our objective is to build a data-driven, actionable, and effective solution for ParTech to help **improve customer engagement and mitigate churn**.

Key-skill Requirements

- Business understanding
- Data pre-processing
- Exploratory data analysis
- Feature engineering
  - Defining target
  - Designing predictive features
- Visualization
- Statistical/Machine Learning techniques, including:
  - Model tuning
  - Model evaluation metrics
  - Model validation
  - Hypothesis testing
  - Experimentation

Process Flow

Exploratory Data Analysis and Feature Engineering

Find relevant insights about the company’s business to build a clear understanding of each feature in the datasets. For this, you need to perform Exploratory Data Analysis and Feature Engineering.

Differentiate between users through Visualizations

Establish a clear definition of customer engagement based on a company's business requirements. This will help you differentiate between engaged and unengaged users. For this, you require relevant business justification, visualizations, and/or modeling techniques.

Predictive Model Building

Design engagement metric that show if customers are likely to be unengaged in the future and the early warnings/drivers of unengagement. For this, you need to predict and classify engaged/unengaged customers using machine learning techniques.

Model Validation and Experimentation

Use the model to identify the most unengaged users. For this, you need to help set up an experiment to prove that the model works.

Dataset Description

Three key datasets are provided by ParTech in the ***enterprise_data*** folder for Data Scientists to explore. These tables consist of information on ~19.3k customers collected over 17 months, at different levels of granularity, depending on the table under consideration. The details of each dataset are as provided below:

Dataset 1: marketing_monthly_data.csv

This table is maintained by ParTech's in-house business team to keep track of CRM activities and customer profiles. It comprises monthly aggregated data on customer demographics, account details, devices, outbound notifications, etc.

Business advice:

- ParTech's internal analytics team often leverages this information to decide the monthly plan of customer outreach or any other form of modeling activities

Feature	Business interpretation
user_id	string uniquely identifying the user
yearmonth	combination of year & month corresponding to observed time-period
city	string corresponding to the user's city of residence
country	string corresponding to the user's country of residence
attributes_notifications_marketing_email	float indicating if the user has accepted to receive marketing push notifications
attributes_notifications_marketing_push	indicating if the user has accepted to receive marketing email notifications
birth_year	integer corresponding to the user's year of birth
user_created_date	datetime corresponding to the user's record creation date
num_contacts	integer corresponding to the number of contacts the user has on the app
num_referrals	integer corresponding to the number of other users referred by each user
num_successful_referrals	integer corresponding to the number of other users successfully (installed and used the app) referred by the selected user
brand_android	integer indicating the user's device type

brand_unknown	
brand_apple	
plan_type_1	integer indicating the plan the user is on
plan_type_2	
plan_type_3	
plan_type_4	
plan_type_5	
plan_type_6	
user_settings_crypto_unlocked	integer indicating if the user has unlocked the cryptocurrencies in the app
channel_push	integer indicating the number of notifications through a particular channel
channel_sms	
channel_email	
reason_reengagement_active_funds	integer indicating the number and purpose of the notifications
reason_pumpkin_payment_notification	
reason_no_initial_card_use	
reason_engagement_split_bill_restaurant	
reason_metal_reserve_plan	
reason_onboarding_tips_activated_users	
reason_made_money_request_not_split_bill	
reason_premium_engagement_inactive_card	
reason_no_initial_card_order	
reason_premium_engagement_fees_saved	
reason_fifth_payment_promo	
reason_welcome_home	
reason_no_initial_free_promopage_card_order	
reason_lost_card_order	

reason_black_friday	
reason_metal_game_start	
reason_joining_anniversary	
status_sent	integer indicating the number and status of the notifications
status_failed	

Dataset 2: transaction_monthly_data.csv

ParTech's in-house analytics team uses the table to comprehend high-level transactional patterns of customers. It comprises of summarized monthly view of customer transactions, including details on value, direction, currency, type, state, etc.

Business advice:

- ParTech's internal analytics teams often use this information to create monthly financial reports for senior business stakeholders or any other form of modelling activities

Feature	Business interpretation
user_id	string uniquely identifying the user
yearmonth	combination of year & month corresponding to observed time-period
amount_usd	float corresponding to the transaction amount in USD
ea_merchant_city	integer corresponding to the number of distinct cities the user has transacted in
ea_merchant_country	integer corresponding to the number of countries the user has transacted in
ea_merchant_mcc	integer corresponding to the number of distinct Merchant Category Codes (MCC's) the

	user has transacted in
transaction_id	integer corresponding to the number of distinct transactions the user has made (this is NOT an ID column)
direction_outbound	integer indicating the number and direction of transactions made by the user
direction_inbound	
ea_cardholderpresence_nan	integer indicating the number of transactions and presence/absence of the cardholder
ea_cardholderpresence_false	
ea_cardholderpresence_true	
ea_cardholderpresence_unknown	
transactions_currency_aed	integer indicating the number and currency of transactions made by the user
transactions_currency_sek	
transactions_currency_aud	
transactions_currency_gbp	
transactions_currency_eth	
transactions_currency_rub	
transactions_currency_chf	
transactions_currency_hrk	
transactions_currency_ltc	
transactions_currency_mad	
transactions_currency_btc	
transactions_currency_nzd	
transactions_currency_jpy	
transactions_currency_ils	
transactions_currency_qar	
transactions_currency_mxn	
transactions_currency_dkk	
transactions_currency_sgd	
transactions_currency_zar	

transactions_currency_bgn	
transactions_currency_usd	
transactions_currency_inr	
transactions_currency_thb	
transactions_currency_ron	
transactions_currency_huf	
transactions_currency_try	
transactions_currency_xrp	
transactions_currency_pln	
transactions_currency_eur	
transactions_currency_bch	
transactions_currency_czk	
transactions_currency_cad	
transactions_currency_nok	
transactions_currency_hkd	
transactions_currency_sar	
transactions_state_completed	integer indicating the number and state of transactions made by the user: **completed** - the transaction was completed, and the user's balance was changed **declined/failed** - the transaction was declined for some reason, usually owing to insufficient balance **reverted** - the associated transaction was completed first but was then rolled back later due to customers reaching out to ParTech
transactions_state_reverted	
transactions_state_declined	
transactions_state_pending	
transactions_state_failed	
transactions_state_cancelled	
transactions_type_transfer	integer indicating the number and type of transactions made by the user
transactions_type_card_payment	
transactions_type_exchange	
transactions_type_atm	
transactions_type_topup	
transactions_type_card_refund	

transactions_type_refund	
transactions_type_fee	
transactions_type_cashback	
transactions_type_tax	
tx_count	integer corresponding to the number of distinct transactions the user has made

Dataset 3: transaction_details_data.csv

This table is owned and analyzed by ParTech's in-house analytics team to understand granular transactional patterns of customers. It encompasses information on each transaction for every customer in terms of datetime, value, state, and type.

Business advice:

- ParTech's internal analytics team uses this information to understand the low-level dynamics of customer transaction behavior and in any other form of modeling activities
- ParTech's internal analytics team advises that this table is likely to have some overlap of information with **transaction_monthly_data.csv** and needs to be used judiciously
- ParTech's business team would prefer a bottom-up approach and advice on leveraging the granular information for deciding:
    - the definition of the target metric which needs to be easy-to-understand, capture data intelligence as well as business value and actionability
    - the optimal window over which the target metric needs to be measured at any point in time (e.g., 3-month future window)

Feature	Business interpretation

transaction_id	string uniquely identifying the transaction (this IS an ID column)
amount_usd	float corresponding to the transaction amount in USD
transactions_type	string indicating the type of the transaction
transactions_state	string indicating the state of a transaction **completed** - the transaction was completed, and the user's balance was changed **declined/failed** - the transaction was declined for some reason, usually owing to insufficient balance **reverted** - the associated transaction was completed first but was then rolled back later due to customers reaching out to ParTech
user_id	string uniquely identifying the user
created_date	datetime corresponding to the transaction's created date

Tasks to be performed

Using the datasets provided, and your analytical knowledge, perform the tasks mentioned below. Each task has been broken down into multiple sub-tasks for the benefit of the Data Scientists.

Task 1:

It is crucial for you as a Data Scientist to gain the confidence of business stakeholders at the very outset of the project through some quick wins in the form of exciting findings, in-depth data understandings, and great visualizations.

As a Data Scientist, you need to dive deep into the datasets (refer to the set of the following sub-tasks) to unlock interesting insights about customers, their transactional behaviours, ParTech's business trends, etc. Also, you need to build a clear understanding of each feature in the datasets through a quick and efficient data quality report.

Sub-tasks - Part A - [Exploratory Data Analysis using Tableau Public]

1. What can be inferred from the distribution of the total **amount_usd** transactions across different **transactions_state**? *(Hint: Use transaction_details_data)*
2. Which **transactions_type** has been most beneficial to ParTech monetarily? *(Hint: Only COMPLETED transactions directly add value to ParTech)*
3. Which type of transactions has the highest **amount_usd** of **non-COMPLETED** transactions?
4. How does the count of new joiners and the total number of transactions by the new joiners vary across time (day-level)? *(Hint: Use user_created_date to identify new joiners)*
5. What does the overall distribution of distinct users across different countries look like on the world map?
6. What does the average **birth_year** of users across different countries look like on the world map?

7. Which plan_type is most popular among the users?
8. How does the average distribution of each **plan_type** vary across yearmonth?
9. What is the total number of SENT and FAILED notifications across yearmonth?
10. What does the box-plot distribution of **amount_usd** and **tx_count** across yearmonth tell us about the outliers?
11. What inferences can be drawn based on average distributions of each transaction **direction** and **transactions_type** across 'yearmonth's?

Sub-tasks - Part B - [Data Quality Report using Jupyter Notebook | Python3]

1. Use pandas-profiling package to auto-create summary reports within the notebook frame, for each of the three datasets

Task 2:

As highlighted in the problem statement, one of the critical areas of focus for ParTech is to understand customer engagement better. The first step towards understanding engagement is to establish a clear and robust definition of the metric. The challenge is, there are no fixed definitions of engagement, and it often needs to be customized based on a company's business requirements.

Through the following sub-tasks, Data Scientists need to build a metric that can help identify engaged users and differentiate between engaged and unengaged users. The approach needs to be backed with necessary business justification, visualizations, rationale, and/or modeling techniques.

Sub-tasks - Part A - [Data Handling using Jupyter Notebook | Python3]

1. Convert the **user_created_date** and **created_date** columns from **marketing_monthly_data** & **transaction_details_data** datasets into usable datetime formats (*Hint: Use pandas package)*
2. Considering 2018, as the observation time-period, what are the steps needed to create a new data frame, such that it satisfies all of the following criteria:
   - Contains 1 record per customer
   - Only COMPLETED transactions are considered
   - Captures average 3-months behavior across the tenure of each customer along four engagement-specific dimensions, as stated below:
     1) Number of transactions *(Use count of amount_usd)*
     2) Total USD value of transactions *(Use sum of amount_usd)*
     3) Total distinct types of transactions *(Use transactions_type)*
     4) Total distinct days of transactions *(Use created_date)*

Hint:

- *2019 data to be used later for testing and experimentation*
- *Tenure = 2018-12-31 23:59:59 UTC - min('user_created_date')*
- *To calculate 3-months average across the tenure of each user, use group-by on user_id for each dimension, divide by tenure in days and then multiply by 90*

Sub-tasks - Part B - [Unsupervised Learning using Jupyter Notebook | Python3]

1. Use two popular clustering algorithms, K-means & Hierarchical, to combine the four engagement-specific dimensions into a single multi-class metric *(Note: Outliers may bias the clustering algorithms)*
2. Compare the centroid values of the two clustering algorithms across each dimension in each cluster
3. Which of the two algorithms is a better choice for defining the engagement metric?
4. How may the centroid values from the optimal algorithm be used to define a business rule for tagging engaged vs. unengaged users?

Note: The above business-rule based engagement metric built on top of unsupervised modeling framework fits ParTech's requirements (easy-to-understand, data intelligence, business value, and actionability-driven engagement metric). However, beyond the scope of this project, Data Scientists are encouraged to research further and explore other robust approaches in designing an engagement metric as well.

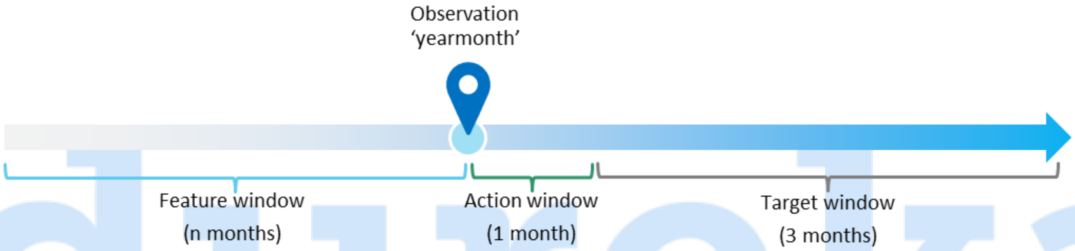
Task 3:

To make the above design of the engagement metric useful to the business, ParTech needs to know which customers are likely to be unengaged in the future. They also need to understand the early warnings/drivers of unengagement so that action can be taken accordingly.

To achieve this, Data Scientists need to predict and classify engaged/unengaged customers using Statistical/Machine Learning techniques, based on the following sub-tasks.

Sub-tasks - Part A - [Model Data Preparation using Jupyter Notebook | Python3]

1. How can the dataset(s) be joined efficiently to create a single dataframe, containing 1 record per customer per yearmonth? *(Note: avoid replication of effort in case of overlap of information across any dataset(s)*)
2. Assume that at any point in time, ParTech needs 1-month heads-up to implement any action (i.e., the gap between feature window and target window needs to be 1-month). Based on this assumption, create a binary engagement target using the engagement metric definition from Task-2)



3. What is the distribution of the target across each yearmonth? Is the target balanced?
4. What are the kinds of feature engineering that can be done using the available information? *(Note: Input features directly correlated with the target metric could lead to overfitting and misinterpretation)*

Sub-tasks - Part B - [Modelling Training and Validation using Jupyter Notebook | Python3]

1. Split the data such that:
   - Training data - 80% till 201809
   - Validation data 1 (out-of-sample) - 20% till 201809
   - Validation data 2 (out-of-time) - 201812
   - Scoring data - 201901
2. Train, tune and validate a Logistic Regression model
3. Train, tune and validate an XGBoost classification model
4. Based on the 2 models, what are the essential indicators of unengagement?
5. Compare the performance of the 2 models. Which model should ParTech use?
6. What can cumulative gains plot tell us about the optimal sample size for testing?

Task 4:

Based on Task-3, ParTech now wants to use the model to identify the most unengaged users and implement some business actions to convert them to engaged users. To enable this, Data Scientists need to help set up an experiment to prove that the model works through the following sub-tasks.

Sub-tasks - Part A - [Model Scoring using Jupyter Notebook | Python3]

1. Use the model developed in Task 3 to score on 201901 data

Sub-tasks - Part B - [Experiment setup using Jupyter Notebook | Python3]

1. Set up test and control groups (60%-40%) from the top 10% unengaged customers. Note: In real-life business scenarios, testing is not usually done on full base but on a smaller set of customers for whom it will be most effective
2. Perform an appropriate statistical test to test customer characteristics and check if control groups are similar

Task 5:

Detailed project report/synopsis.

How to submit your project?

Following are the tasks, which need to be developed while executing the project:

- Share your project via google colab to *support@edureka.co*
- The .ipynb file with details of each step in the markdown
- You can even upload your code into your github repository and share your repository with us

Marks Allocation

Task 1:

- Part A [11 x 2 Marks]
- Part B [1 x 3 Marks]

Task 2:

- Part A [2 x 4 Marks]
- Part B [4 x 5 Marks]

Task 3:

- Part A [2 + 4 + 2 + 4 Marks]
- Part B [2 + 4 + 4 + 2 + 2 + 2 Marks]

Task 4:

- Part A [1 x 3 Marks]
- Part B [2 x 3 Marks]

Task 5:

- Project report/synopsis with detailed .ipynb (with best Markdown explanation) [10 Marks]