BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

Work Integrated Learning Programmes Division

M.Tech.in Data Science and Engineering

Prediction of Produt upgrade of freemium/trial users

DISSERTATION REPORT
DSECLZG628T

Submitted in partial fulfillment of the requirements of the

MTech Data Science and Engineering Degree programme

By

Prasanth R
2019hc04979

Dissertation work carried out at
Trimble Information Technologies India Pvt Ltd, Chennai

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE
Pilani (Rajasthan) INDIA

(February 2022)

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

Work Integrated Learning Programmes Division

M.Tech.in Data Science and Engineering

Prediction of Produt upgrade of freemium/trial users

DISSERTATION REPORT
DSECLZG628T

Submitted in partial fulfillment of the requirements of the

MTech Data Science and Engineering Degree programme

By

Prasanth R
2019hc04979

Under the supervision of

Sabari Murugan S,
Senior Software Engg Lead,
Trimble Information India Pvt Ltd.,

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE
Pilani (Rajasthan) INDIA

(February 2022)

ACKNOWLEDGEMENT

Foremost, I would like to express my sinere gratitude to my mentor **Sabari Murugan S**, for the continuous support during this entire M.Tech course and the final year dissertation. His patience, motivation, enthusiasm, immense knowledge and guidance helped me a lot in the whole course tenure. I could not have imagined having a better advisor and mentor for the course.

I am extremely thankful and pay my gratitude to my faculty **Govada Aruna** for her valuable guidance and support on the completion of this project in its presently.

Secondly, I would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

Prasanth Rajendran

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

CERTIFICATE

This is to certify that the Dissertation entitled **Prediction of Product upgrade of freemium/trial users** and submitted by Mr.**Prasanth R** ID No.**2019HC04979** in fulfillment of the requirements of **DSECLZG628T** Dissertation, embodies the work done by him under my supervision.

S. Sa

Place: Chennai
Date: 20 Feb 2022

(Signature of Supervisor)
Sabari Murugan S
Senior Software Engg Lead

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
SECOND SEMESTER 2020-21

DSECLZG628T DISSERTATION

Dissertation Title : **Prediction of Product upgrade of freemium/trial users**

Name of Supervisor : **Sabari Murugan S**

Name of Student : **Prasanth R**

ID No. of Student : **2019hc04979**

Abstract

Every SaaS product companies have freemium/trial-based product offerings to let the user evaluate their product's worthiness provided with limited features. Every company's motive is to upgrade the trial user to a premium user, however, in reality, a considerable percentage of product evaluators use the product with limited features till the trial period or forever. Also, users who leverage the portal effectively during the trial tenure are considered to be potential candidates for premium subscribers. So reaching out to the pertinent consumer is a key growth pillar for the product's revenue which will also help to focus adverts towards certain customer groups of an audience instead of reaching to all the consumers who use freemium/trial features.

Analyzing the active trial users based on their usage pattern of the free product during the trial period is critical for projecting the product revenue. The analysis is generally carried out by factoring in the number of times the user logged in, the time spent actively in the various product screens and the features explored, and users' interest in premium upgrade option inquiry such as exploring the price options.

As part of this dissertation, models using prescriptive techniques i.e., Uplift modeling will be trained based on the prepared dataset to simulate multiple possible scenarios of user activities and usage patterns. Once a model is trained it will be used to predict whether a user will opt for the premium features or not.

(Signature of Student)

Prasanth R

2019hc04979

(27/November/2021)

(Signature of Supervisor)

Sabari Murugan S

Senior Software Engg Lead

(27/November/2021)

5. List of Symbols & Abbreviations used

1. AWS - Amazon Web Service
2. EDA - Exploratory Data Analysis
3. SaaS - Software as a service
4. REST API - Representational State Transfer
5. IDE - Integrated Development Environment

6. List of Tables

7. List of Figures

Table of Contents

8. Dissertation Details

8.1. Introduction and Background

Trimble Connect, a SaaS product facilitates the trial usage called Freemium which helps the customers to evaluate the product with certain limitations. In general terms Freemium is different from Limited period access where the later allow the users to experiment the product only during the certain period of timeline for instance, 3 months and the prior allow the users to try the features forever with limited previledges. Google drive is a typical SasS product example for Freemium product which provides the infinite timeline access with certain limitations in storage and features.

Knowing the customer pattern is crucial part in this business to upkeep the success of a product. As per the Stats only 25 to 30 percentage of the users who upgrade their license to Premium, hence the efforts of targeting the remaining 70 with ads and other marketing programs are not a potential opportunities for revenue generation which may also result in redundant expenditure for the organizations. Moreover finding out the 30 percent of active users who experiment the product effectively paves a way narrow down the Sales initiatives. Hence, developing the Machine Learning module to identify the potential candidates is the need of the hour solution which helps the organization to followup with active users to purchase the premium license by providing whooping discounts, benefits and free trainings.

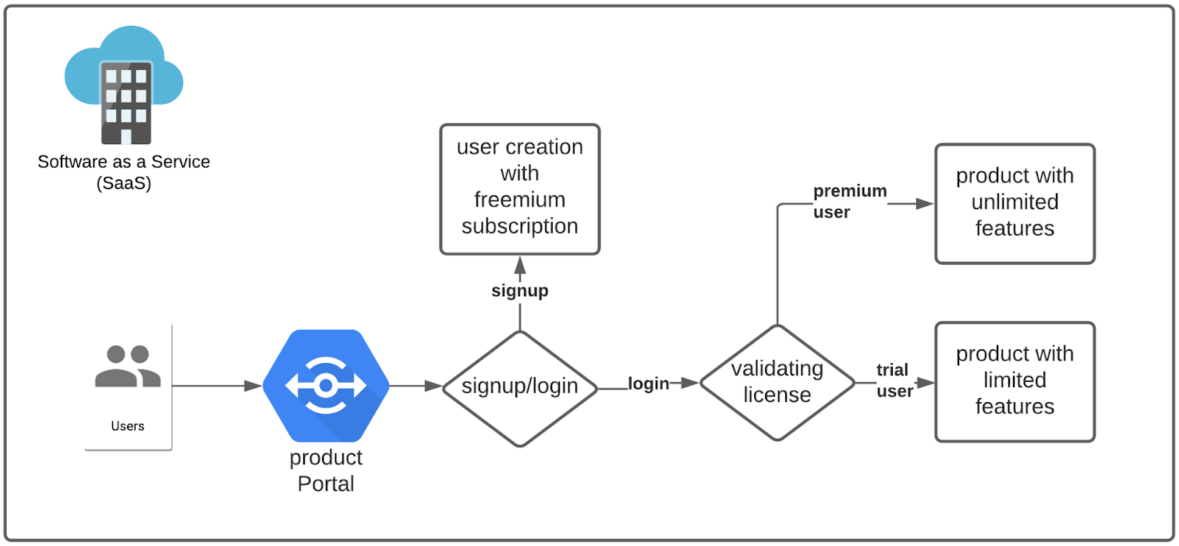
Nevertheless, defining the active users characteristics and behaviour patterns are the rudimentary steps to train a model, which involves tracking the specific patterns of every users' usage. The analysis is generally carried out based on the factors like the number of times the user logged in, the time spent actively in the various product screens and the features explored, and users' interest in premium upgrade option inquiry such as exploring the price options.

The project that we develop as part of this dissertation monitors and trains the data collected from the usage of application by freemium users and predicts the potential users who are likely to purchase the Business license. Using this prediction the users can be given personalized recommendations and offers by sharing information about the perks of being a business licensed user. This will help the Trimble Connect product team as well as marketing team to target the customers in a more efficient way there by improving the revenue on a higher scale.

8.2. Problem statement

As a Saas products provide Freemium features with limited access for the users to explore the product, there is a necessity to know the user's behavior to predict whether the product subscription will be upgraded or not, so developing such prediction models may provide some insights to reach the potential upgradable customer which in turn benefits the product in terms of revenue. Hence the problem statement is to predict whether SaaS product users with freemium/trial-based licenses will likely purchase/upgrade the premium version based on the user's usage pattern.

8.3. Business Flow



Software as a Service (SaaS)
user creation with freemium subscription
premium user
product with unlimited features
signup
Users
signup/login
login
validating license
trial user
product with limited features
product Portal

Figure 1

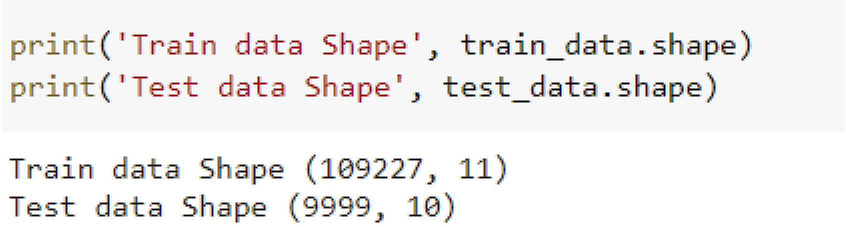
8.4. Objective of the project

8.4.1. Data Collection

It is an important step to start with the dissertation. We have collected a real time user data without compromising the data quality at the same time without breaching any data integrity or involving in any form of data violations. The trivial user related(will not be useful for predicts) are excluded and some are mocked.

We have created an application in Java to collect the realtime data. The collected raw data is not gathered simply by querying the tables. There are umpteen transformations and aggregations are performed to collect the reliable data which provides the pertinent results.

We have collected training data as well as the test based on the real data as part of this dissertation. Please refer to the below details for the collected data.

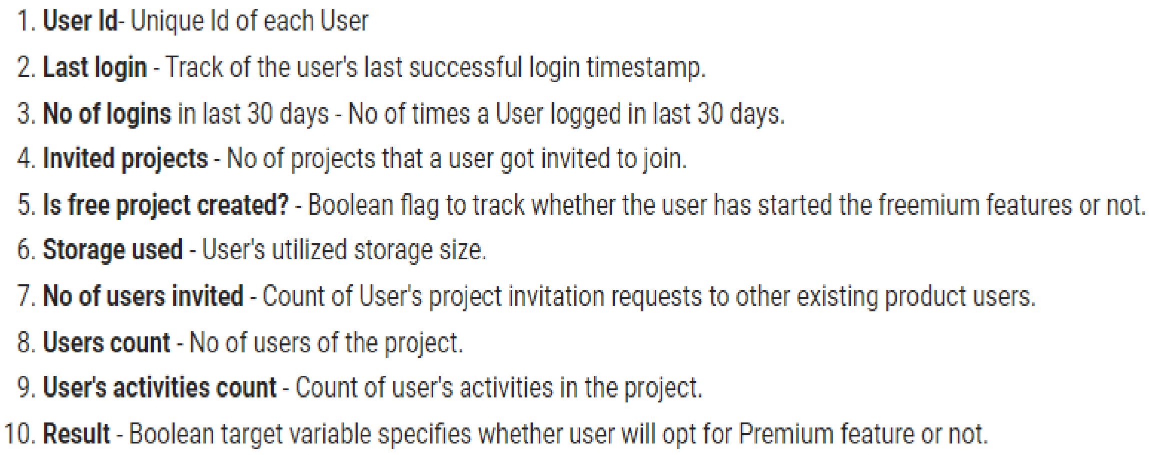


```
print('Train data Shape', train_data.shape)
print('Test data Shape', test_data.shape)
```

```
Train data Shape (109227, 11)
Test data Shape (9999, 10)
```

Figure 2

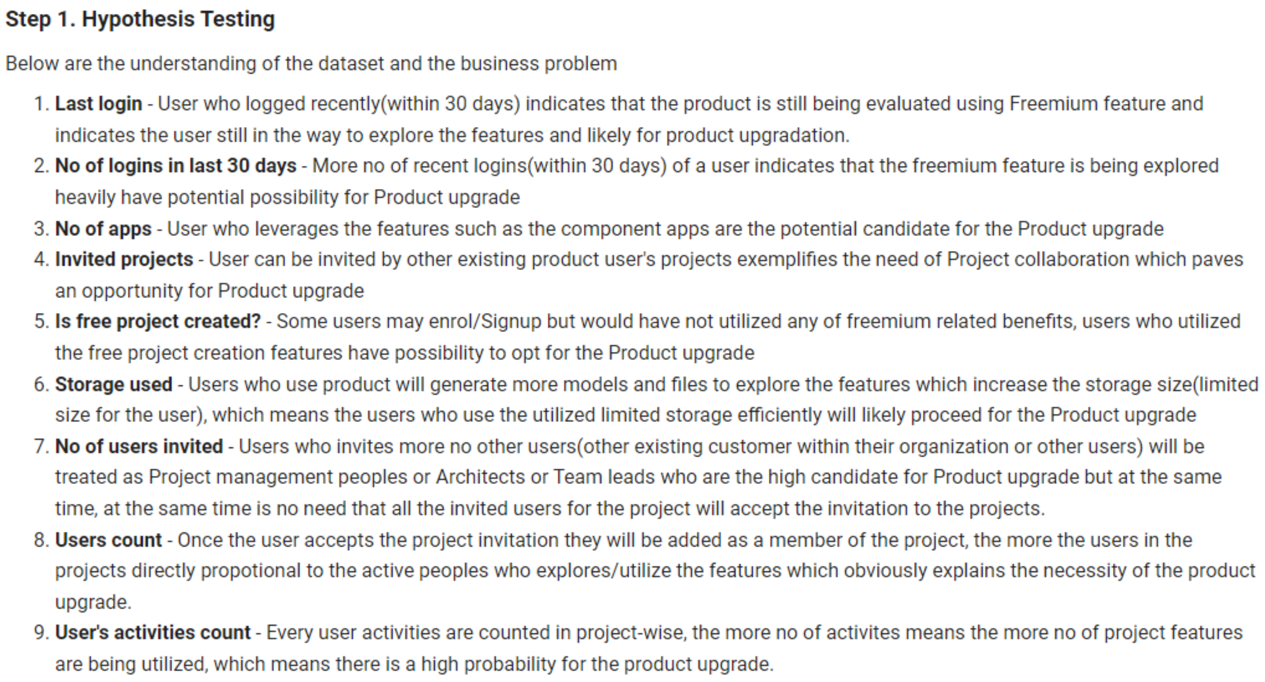
Details about the the variables gathered in our dataset.



1. **User Id**- Unique Id of each User
2. **Last login** - Track of the user's last successful login timestamp.
3. **No of logins** in last 30 days - No of times a User logged in last 30 days.
4. **Invited projects** - No of projects that a user got invited to join.
5. **Is free project created?** - Boolean flag to track whether the user has started the freemium features or not.
6. **Storage used** - User's utilized storage size.
7. **No of users invited** - Count of User's project invitation requests to other existing product users.
8. **Users count** - No of users of the project.
9. **User's activities count** - Count of user's activities in the project.
10. **Result** - Boolean target variable specifies whether user will opt for Premium feature or not.

Figure 3

8.4.2. Feature Engineering and Feature Selection



Step 1. Hypothesis Testing

Below are the understanding of the dataset and the business problem

1. **Last login** - User who logged recently(within 30 days) indicates that the product is still being evaluated using Freemium feature and indicates the user still in the way to explore the features and likely for product upgradation.
2. **No of logins in last 30 days** - More no of recent logins(within 30 days) of a user indicates that the freemium feature is being explored heavily have potential possibility for Product upgrade
3. **No of apps** - User who leverages the features such as the component apps are the potential candidate for the Product upgrade
4. **Invited projects** - User can be invited by other existing product user's projects exemplifies the need of Project collaboration which paves an opportunity for Product upgrade
5. **Is free project created?** - Some users may enrol/Signup but would have not utilized any of freemium related benefits, users who utilized the free project creation features have possibility to opt for the Product upgrade
6. **Storage used** - Users who use product will generate more models and files to explore the features which increase the storage size(limited size for the user), which means the users who use the utilized limited storage efficiently will likely proceed for the Product upgrade
7. **No of users invited** - Users who invites more no other users(other existing customer within their organization or other users) will be treated as Project management peoples or Architects or Team leads who are the high candidate for Product upgrade but at the same time, at the same time is no need that all the invited users for the project will accept the invitation to the projects.
8. **Users count** - Once the user accepts the project invitation they will be added as a member of the project, the more the users in the projects directly propotional to the active peoples who explores/utilize the features which obviously explains the necessity of the product upgrade.
9. **User's activities count** - Every user activities are counted in project-wise, the more no of activites means the more no of project features are being utilized, which means there is a high probability for the product upgrade.

Figure 4

Below task are performed as part of the feature engineering

- Handled the missing values, outliers, unbalanced dataset
- Data transformation of categorical values to numerical values
- Created required features from the existing feature set by applying use case logics
- Cleaned the data by dropping unique value, temporal features which helps to create meaningful features
- Feature selection is done on top of scalar transformed data and the features are finalized using the ‘feature_selection’ library with feat_importances score

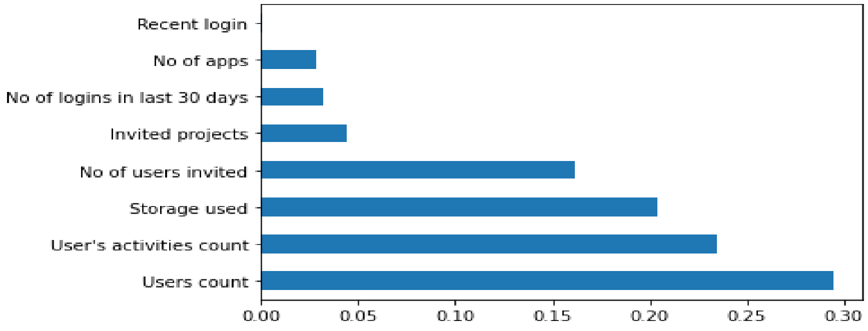


Figure 5

- Selecting a high accuracy classification algorithm by comparing multiple classification algorithms.

8.4.3. Model Building, Deployment and Predictions

The model is build on top of the finalized features using various techniques.

As the problem statement belongs to classification model related, we started the model building process with the basic “LogisticRegression” algorithm and build the “DecisionTree”, “RandomForest” and “knn” models using the processed dataset.

We compared the classification algorithms based on ROC curve and accuracy score also using k-fold Cross-Validation. Based on the aforementioned result “RandomForest” performed better than other algorithms. Traditional Propensity Ensemble methods for uplift modeling such as “XGBClassifier” will be used to increase the accuracy of the model and for final predictions.

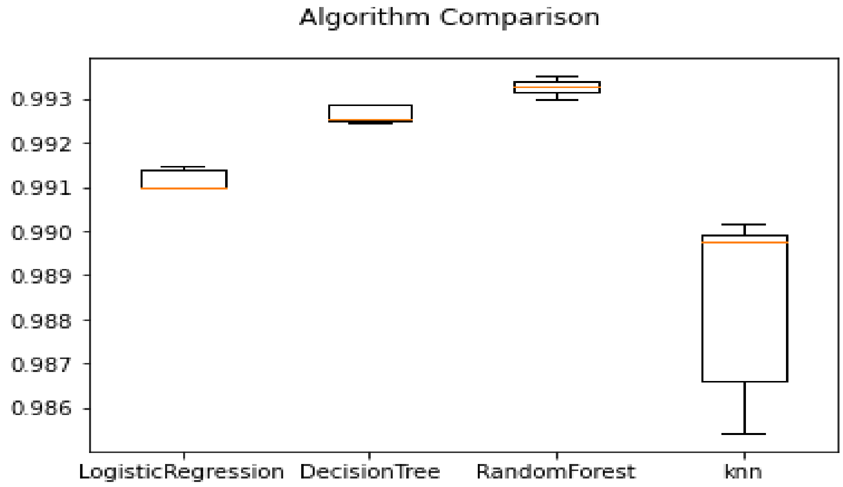


Figure 6

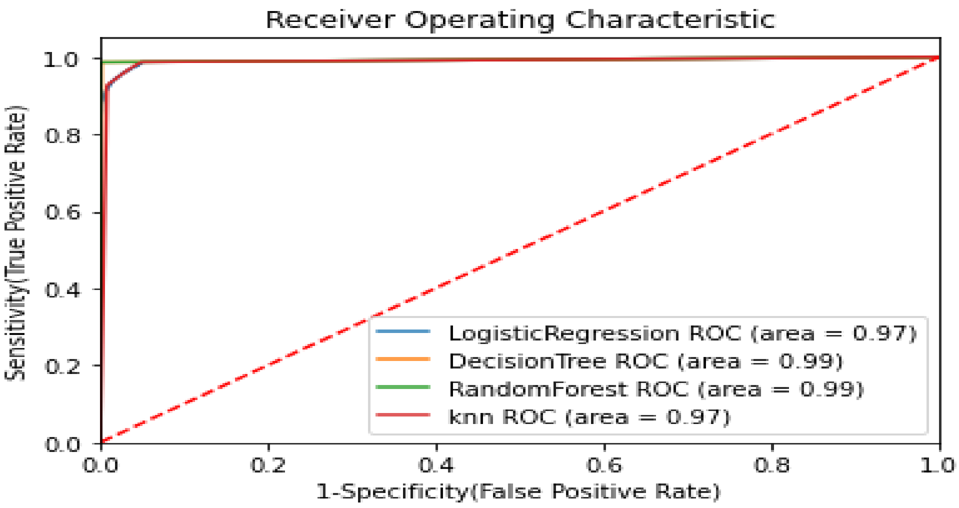


Figure 7

8.5. Uniqueness of the project

This project effort helps to narrow down the potential premium subscribers of a Saas product using prescriptive techniques i.e, Direct Uplift Models. This project monitors and trains the data collected from the usage of application by FREEMIUM users and predicts the potential users who are likely to purchase the Business license.

8.6. Benefit to the organization

Working as one of the core developers in a Saas product, this project will help organizations to find potential premium subscribers and also help them to target focused advertisement instead of wasting resources in an advertisement for non-potential users

8.7. Scope of work

After the model is built and trained, it is deployed to AWS SageMaker, the AWS SageMaker provides model endpoint features using Amazon API Gateway and AWS Lambda which helps us to make API call based prediction. We have developed python Flask based app to access the prediction results through REST APIs.

8.8. Resources needed for the project

- Machine learning python libraries
- Google Colab
- IDE/Jupyter notebook
- Custom developed Java application for data collection
- AWS SageMaker
- Postman API client tool

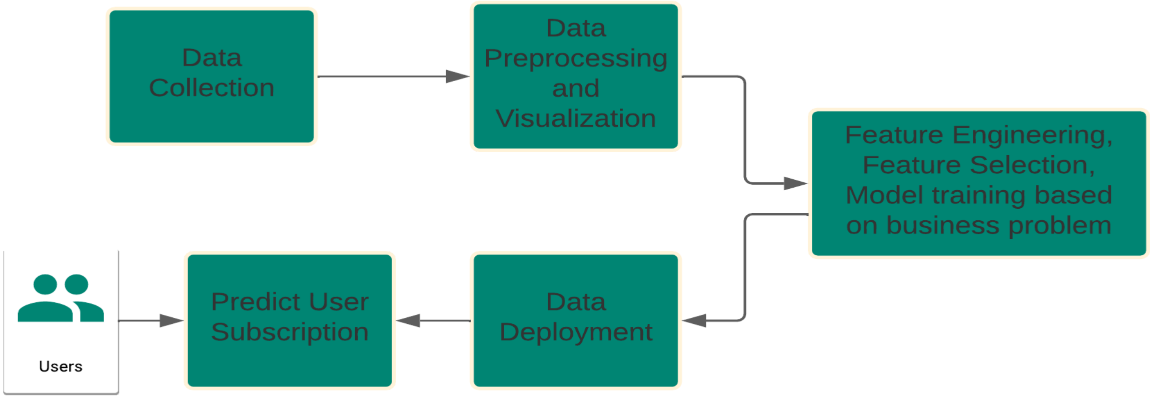
8.9. Potential challenges & risks in doing the project

Collecting the data for this project was really challenging, which involves developing an Java application to collect potential non-trivial raw data from various data sources and performing the aggregation as the raw data mostly contains traditional row based data for every user's activity. As part of the data accumulation process, we have written various complex aggregation logics to streamline the data from various resources. The data should not be collected with user information which violates the data regulations, hence mocking the trivial user data(will not be useful for predictions) without compromising the data quality for data analysis.

8.10. Background of previous work done in the chosen area

There are several efforts/development/research happening in the areas of Predicting User churn, Customer retention, Customer satisfaction and uplift modeling.

8.11. Solution architecture



Data Collection
Data Preprocessing and Visualization
Feature Engineering, Feature Selection, Model training based on business problem
Data Deployment
Predict User Subscription
Users

Figure 8

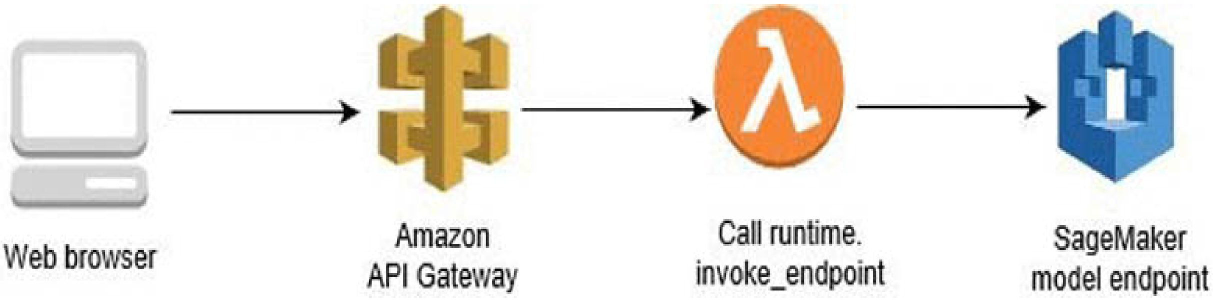


Figure 9

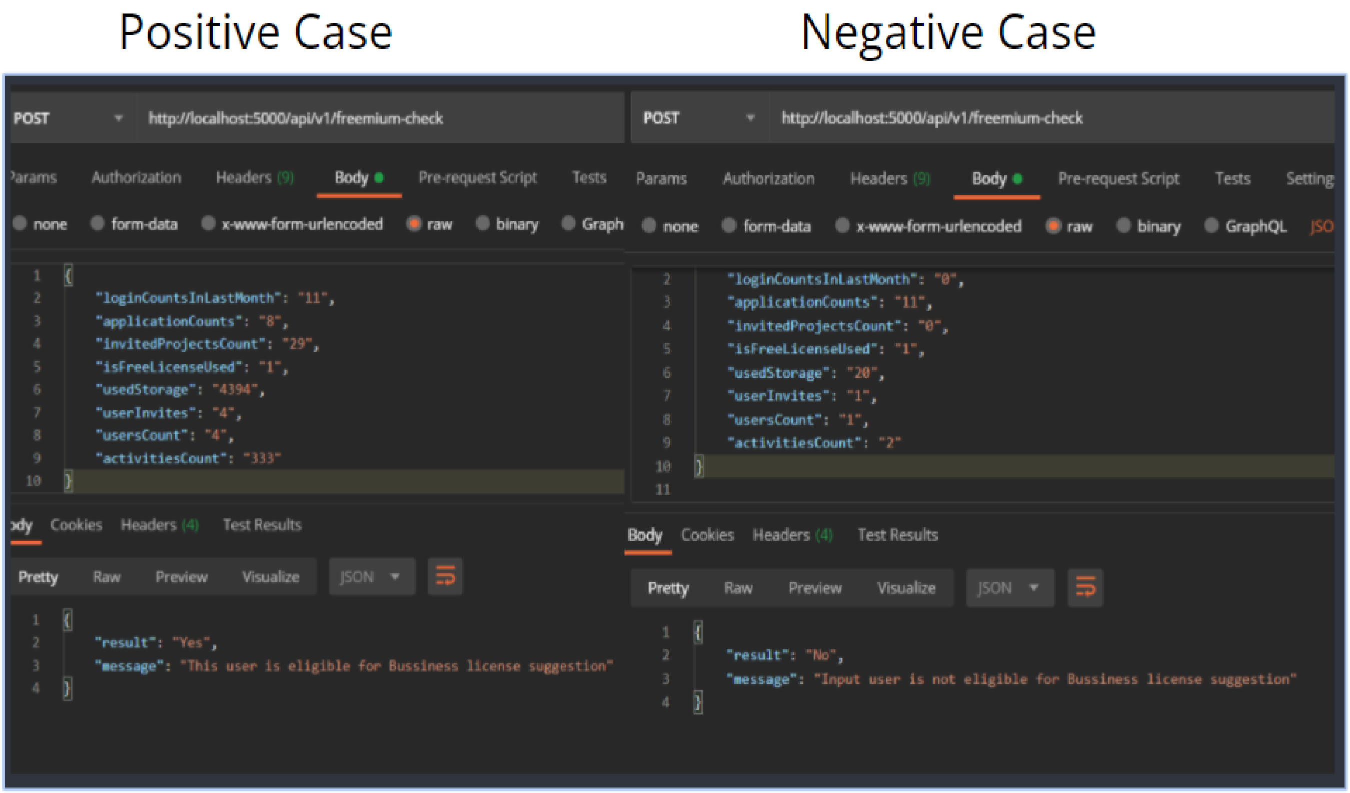
8.12. Project Plan & Deliverables

#	Task	Expected date of completion	Names of Deliverables
1	Data Collection and Understanding – EDA Phase 1	Dec 2nd Week	Raw DataSet
2	Data cleanup/ transformation/ preprocessing – EDA Phase 2	January 2week	Refined Dataset to progress and completion of EDA
3	Model building/train and test dataset including feature engineer	Feb 2nd week	Algorithm Selection including the model building completion
4	Final prediction user model subscription flow	Feb 4th week	Forecast Prediction of User Freemium to Premium conversion
5	Final project report	Feb 28	Dissertation final report submission
6	Presentation	March 2nd Week	Viva demo

Table 1

9. Conclusions

As per the scope of the work, we have evaluated many classification algorithms “LogisticRegression”, “DecisionTree”, “RandomForest” and “knn” and compared them with accuracy score using k-fold Cross-Validation. We also evaluated traditional propensity ensemble methods for uplift modeling such as “XGBClassifier” to increase the accuracy further. After the model is built and trained, it is deployed to AWS SageMaker. We have developed python Flask based application to access the prediction results through REST APIS which is achieved through the combination of AWS SageMaker model endpoint features with Amazon API Gateway and AWS Lambda. Now the predictions can be access using the APIS which helps us yield the target predictions based on the requested details. The below are the screenshot exemplifies, the predictions are being accessed using Rest API calls



Positive Case

POST http://localhost:5000/api/v1/freemium-check

```
{
    "loginCountsInLastMonth": "11",
    "applicationCounts": "8",
    "invitedProjectsCount": "29",
    "isFreeLicenseUsed": "1",
    "usedStorage": "4394",
    "userInvites": "4",
    "usersCount": "4",
    "activitiesCount": "333"
}
```

```
{
    "result": "Yes",
    "message": "This user is eligible for Bussiness license suggestion"
}
```

Negative Case

POST http://localhost:5000/api/v1/freemium-check

```
    "loginCountsInLastMonth": "0",
    "applicationCounts": "11",
    "invitedProjectsCount": "0",
    "isFreeLicenseUsed": "1",
    "usedStorage": "20",
    "userInvites": "1",
    "usersCount": "1",
    "activitiesCount": "2"
}
```

```
{
    "result": "No",
    "message": "Input user is not eligible for Bussiness license suggestion"
}
```

Figure 10

10. Directions for future work

Improving the prediction algorithm to suggest the appropriate license to the user based on their usage. Enhancing the prediction technique to suggest different features for the users.

11. References

https://towardsdatascience.com/product-analytics-drives-freemium-conversions-in-mobile-gaming-and-beyond-baa9b0ef139d

https://www.absolutdata.com/blog/freemium-to-premium/

https://tomtunguz.com/data-science-in-freemium-businesses/

https://getthematic.com/insights/5-ways-data-and-text-analytics-improve-customer-retention/

https://towardsdatascience.com/uplift-modeling-e38f96b1ef60

https://scholarspace.manoa.hawaii.edu/bitstream/10125/50002/1/paper0115.pdf

12. Appendices

1. Trimble Connect

   Trimble Connect is a cloud-based collaboration platform with real time status-sharing.
2. Amazon Web Service

   Amazon Web Services is a on-demand cloud computing platforms
3. Postman API client

   Postman is an API platform for building and using API

13. Duly Completed Checklist

Check list of items for the Final report

a) Is the Cover page in proper format? Y
b) Is the Title page in proper format? Y
c) Is the Certificate from the Supervisor in proper format? Has it been signed? Y
d) Is Abstract included in the Report? Is it properly written? Y
e) Does the Table of Contents page include chapter page numbers? Y
f) Does the Report contain a summary of the literature survey? Y

- Are the Pages numbered properly? Y
- Are the Figures numbered properly? Y
- Are the Tables numbered properly? Y
- Are the Captions for the Figures and Tables proper? Y
- Are the Appendices numbered? Y

g) Does the Report have Conclusion / Recommendations of the work? Y
h) Are References/Bibliography given in the Report? Y
i) Have the References been cited in the Report? Y
j) Is the citation of References / Bibliography in proper format? Y

(Signature of Student)

Prasanth R

2019hc04979

(26/Feb/2022)

(Signature of Supervisor)

Sabari Murugan S

Senior Software Engg Lead

(26/Feb/2022)