2021-2022 Term 1 FTEC5510
Adanced Financial Infrastructure

Group 6

Personalized E-commerce Product Recommendations

3 December 2021

Tutor

Prof. CHAN, Chun Kwong

Group members

LIU Hao Fei (1155164882)

RAO Jia Yi (1155164916)

YIM Ka Wai (1155160891)

YU Tian Yi (1155164881)

Abstract

In this report, it illustrates the overview of virtual bank development in Hong Kong and the difficulty virtual bank are facing currently. A recommendation modeling and big data analytical reports will be introduced to provide them with an insight into their customers and their product preference. It is expected that based on the above outputs, virtual banks can understand their customer, provide personalized marketing to acquire and retain the customer and eventually enhance their market penetration in the banking industry.

In the following sections, it will explain the solution in detail, including but not limited to high-level process flow, system architecture, modeling procedure, data visualization.

Section 1. Introduction

1.1. Background

With the rapid financial technology advancement, the bank industry landscape has been experiencing substantial changes. The Hong Kong Monetary Authority (HKMA) introduced a package of initiatives in 2017 to bring Hong Kong into a new era of digital banking [1]. One of them was the introduction of virtual banks. Virtual bank is a neobank which provides services to the retail segment and small and medium enterprises (SMEs) via electronic channels rather than physical bank branches. This is expected to offer an entirely new customer experience, promote financial inclusion, and enhance innovation in Hong Kong.

1.2. Motivation

Although HKMA granted licenses to 8 virtual banks and virtual banking has been becoming increasingly popular in Hong Kong [4], it encountered intense market competition in the banking industry. Per Finder's survey in Oct 2021 [7], only 18% of respondents had a virtual-only bank account. Major traditional banks such as The Hongkong and Shanghai Banking Corporation (HSBC) continue to dominate the local market share [5]. Therefore, this project aims to investigate how to accelerate customer acquisition and retention in digital banking by improving customer experience using artificial intelligence (AI) and big data. It will then help virtual banks to deliver a highly customized cyber experience for their clients, differentiate themselves from traditional banks and drive rapid market penetration in Hong Kong.

1.3. Innovative Element

In this project, a recommendation system with analytical report is created to provide virtual banks a financial technology solution. The technical solution is as follows:

1. **Customer Segmentation:** Dividing customers into several groups based on the similarity between customers and their historical transaction
2. **Personalized Cross-selling Recommendation System:** Analyzing data of each customer group and providing the most suitable financial services.
3. **Tailor-made Report:** Creating advanced analytical reports using big data approach

1.4. Experimentation

The process consists of five steps:

1. **Research:** Understanding various predictive modeling approaches and determining the optimal solution to be used to build the recommendation system on customer behavior data.
2. **Hypothesis:** Expecting the performance of machine learning algorithm (i.e., Personalized recommendation system) will provide suitable services to each customer group
3. **Machine learning**: Building models for customer segmentation and recommendation system and tuning the model hyperparameters to enhance the model performance
4. **Performance evaluation**: Comparing the accuracy of different models and selecting the most appropriate models to be delivered on the web page
5. **Development**: Creating tailor-made customer reports and personalized recommendation system

Section 2. Research

2.1. Virtual Bank in Hong Kong

2.1.1. Background

Virtual bank is defined as a bank which primarily delivers retail banking services through the internet or other forms of electronic channels instead of physical branches [2].

According to HKMA [1], the introduction of virtual banks in Hong Kong is a key pillar supporting Hong Kong's entry into the Smart Banking Era. The HKMA believes that the development of virtual banks will promote fintech and innovation in Hong Kong and offer a new kind of customer experience. In addition, virtual banks can help promote financial inclusion as they normally target the retail segment, including the small and medium-sized enterprises (SMEs).

In Sep.2017, HKMA launched an array of initiatives under the title of “A New Era of Smart Banking” and introduced the “Virtual Bank” license [4]. Recognizing the disruptive potential of providing an operationally streamlined, digitally native banking proposition, 29 aspirants submitted application to the HKMA for a virtual banking license. Following a rigorous selection process, eight applicants were granted approval for a license in 2019, all of which have since launched their operations. Along with the vision of the HKMA, in 2020, Hong Kong hit a major Fintech milestone with eight virtual banks fully launched. They are ZA Bank, Airstar Bank, Welab Bank, Livi Bank, Ant Bank, Mox Bank and Fusion Bank [3]. The procedure graph of launching time and the list of virtual banks are posted in ***Appendix 10.4.1 and Appendix 10.4.2.***

2.1.2. Current Development

After receiving their licenses in 2019, Hong Kong’s eight virtual banks began actions in 2020, but it was lack of influence due to the Covid-19 pandemic [3]. While most of these banks have been operating for less than a year, their latest financial results provide an indication of their initial progress and growth prospects going forward. However, all the virtual banks are yet to turn around a profit as they remain focused on investing and spending on information technology and marketing activities to grow their customer base. We can see the losses before tax of these eight banks (see ***Appendix 10.4.3***).

Combined total deposits of all the virtual banks as of December 2020 was around HK$15.8 billion but is just represent 0.11% of total deposits across the entire banking sector [3]. ZA Bank and Mox Bank had the largest share of deposits at 38% and 33% respectively (see ***Appendix 10.4.4***).

Authorized financial institutions in Hong Kong hold HKD 14.6 trillion worth of customer deposits, with HKD 10.5 trillion circulating in the form of loans and advances [12]. While the city is home to 2,300 lenders, the majority of Hong Kong’s banking industry has been cornered by 172 licensed banks [13], with a select group of legacy institutions controlling the lion’s share of various lending activities. We estimate that the top four banks – HSBC, BOCHK, Hang Seng Bank, and Standard Chartered, accounting for 62% of total deposits and 54% of the total lending market (see ***Appendix 10.4.5***).

When it comes to Hong Kong banking and finance app industry, it is showed that traditional banking apps ranked Top 20 in the market, while virtual banking apps had comparatively lower penetration – ZA Bank ranked 30th (see ***Appendix 10.4.6***). Virtual banks have been steadily gaining momentum since 2020. Hong Kong has long been recognized as a global financial hub. Supported by its free internet access and high smartphone penetration, digital banking services continue to gain popularity, which is also reflected by the increasing penetration of banking & finance-related apps. Certainly, it is inescapable that fierce competition would happen among these eight virtual banks. ZA Bank is the first virtual bank officially launched in Hong Kong; it still maintains its competitive position with the highest penetration in the market. However, it's also worth looking at the similar penetration from 2nd to 5th position, prompting fierce competition between these virtual banks, on aggressively building their customer base.

Virtual bank, as a new market player, its overall penetration is still lower than traditional banking apps, like HSBC and Hang Seng Bank, which have a long history of establishment and strong market dominance.

Therefore, virtual banking still has a long way to go since traditional banking apps are still the most ubiquitous in Hong Kong.

2.1.3. Marketing

When Hong Kong's virtual banking apps appear, they usually had some attractive publicity and promotional campaigns, which helped them build market presence and boost the number of app users (see ***Appendix 10.4.7***). For instance, ZA acted in concert with Cash Payout Scheme to launch "Why get $10,000 when you can get $11,000?" and the contest is organized by TVB and sponsored by Fusion Bank with digital red packets and cash coupons worth over 20 million.

However, although they built brand presence is the first step, they were not able to maintain a sustainable install rate and retain loyal users. So, it is a good idea to utilize their big data and technologies to understand your current customers and uncover your potential customers.

2.1.4. Banking Service

In a fiercely competitive market like Hong Kong, it is essential for virtual banks to continue to expand their products and services to stand out from their peers and gain a critical mass of customers and deposits. Keen to deliver on their promise of revolutionizing the banking industry, the virtual banks have made an immediate mark on the city's banking industry by [5]:

(1) enabling rapid account opening within a few minutes.
(2) waiving many fees and minimum balance requirements.
(3) providing remote banking services through digital means.
(4) offering relatively higher interest rates on savings accounts and time deposits.
(5) extending speedy monetary / payment transfer capabilities.
(6) delivering tailor-made personalized offerings.
(7) providing 24/7 customer service through digital channels.

At present, Hong Kong's virtual banks have launched relatively basic products and services, such as savings accounts and payment capabilities, with several players also offering time deposits and debit cards. Currently, ZA Bank and Mox Bank appear to have the widest range of offerings, with the former being granted an insurance agency license by the Hong Kong Insurance Authority ("IA") to extend insurance offerings, and the latter offering innovative budgeting tools (see ***Appendix 10.4.8***).

In addition to providing highly competitive lending and deposit rates, the virtual banks are offering free of charge digital and QR-code payments and transfer capabilities, along with free-to-use budgeting tools. Like many startups, the provision of no cost services highlights a clear priority for the eight new payers: rapidly scale their customer base by attracting consumers looking for low-cost financial services offerings.

2.2. Predictive Modeling Approach

According to the international journal of engineering research and technology [6], Recommendation system plays a critical role in business. It can be treated as marketing strategy by suggesting appropriate product/service to the user based on the historical data. Customer satisfaction can then be achieved following with customer loyalty. Different techniques can be used for filtering. It can be categorized into three types. The advantage and disadvantage of each technique is summarized in ***Appendix 10.4.9.***

First, it is collaborative filtering. It consolidates and analyzes a large amount of data and predicts what users will prefer based on the similarities between users, so called people-to-people correlation. This filtering can be further split into two types – memory-based method and model-based method [9]. Memory-based method makes recommendations by using database directly, so it allows data change, but it needs large computational time. Model-based method creates a model using the transaction data to perform recommendation. Itt has smaller computing time as it is not affected by data size, but data change is not allowed. Collaborative filtering is the most widely used mainly due to the ability to precisely recommend serendipitous and/or complex items without dependency of machine-readable representation of the items to be recommended. However, collaborative filtering approach often suffers from several limitations, for example, cold start, scalability, sparsity.

Second, it is content-based filtering. It uses the similarities among product/service or content feature, so called item-to-item correlation, and recommends the items that are related to the items a user liked in the past. The recommendation is specific to a particular user, so it does not require data about other users, and it is easier to scale to a huge number of users. Also, it can capture specific interest of a user and suggest items not many users are interested in. Nevertheless, content-based filtering is restricted to recommend the product only according to users' existing interests. Even including the descriptive data, some experiments found that collaborative filtering will provide more accurate results than content-based filtering.

Third, it is demographic filtering. It classifies the user into different demographic classes based on personal attributes, finds the range of products each class is interested in and recommends product/service accordingly. This approach does not need large and complex data collection, such as purchasing history of users, but the above classification could not be specific enough and it will lose the users' individuality. Moreover, personal information is provided by customers. It is possible that the data is incomplete and/or incorrect. And it does not support a changed user profile in classification.

2.3. Dataset Split Ratio

To build a predictive modeling, the dataset is required to split into two subsets, which are training set and variation set. Training set is the data sample used to fit the model while variation set is the data sample used to evaluate the model fit on the training set and fine-tune the model hyperparameters. According to Tarang Shah [10], the dataset is generally split into training set and validation set with an 80:20 ratio Although there is no optimal dataset split ratio and it is specific to the use case, it depends on the following two factors, which are the number of data sample in total and actual model to be trained. When the model training requires substantial data, a larger training set is needed. If the model has very few hyperparameters, it is easier to validate and tune so the validation set can be reduced, vice versa. After the model is finalized, another set of data, so called testing set will be used to perform unbiased evaluation on the model performance.

Section 3. Methodology

3.1. Methodology and Tool

A financial product recommendation system is built via Python. It involves hybrid filtering technique (memory-based collaborative filtering and demographic filtering). A hybrid filtering approach is used because memory-based collaborative filtering can discover new interests for each user without domain knowledge. However, it has a critical drawback, which is cold start. It means that collaborative filtering is not able to deliver recommendation for a new virtual bank customer. Hence, demographic filtering is also introduced into the modeling so virtual bank can still use customer personal information to provide personalized marketing and increase its profitability.

There are two types of customer related report to be given to virtual banks: overview report and detailed report. For overview report, it aims to give virtual bank a broad insight on its customer. Data visualization aims to express information clearly and effectively by means of graphics. Business Intelligence Visualization tools can be used to help derive value from data. In this project, R language (ggplot) and Python (Matplotlib, Seaborn, Plotly) are used for data visualization since it can gather a large amount of data and provide different data visualization efficiently. For detailed report, it aims to provide every customer's product preference and product mapping so that virtual bank can provide its specific product/service to each customer accordingly. As it involves the product mapping, it is hard to automate the reporting process in this stage. Therefore, this detailed report is manually generated based on our business analysis.

3.2. Customer Process Cycle

Customer process cycle is a high-level graph illustrating how the recommendation system and analytical report can be used by virtual bank to enhance customer acquisition and retention (see ***Appendix 10.5***)***.*** The process flow consists of three sections: acquisition, monetization and maintenance.

In the acquisition section, virtual bank collects the customer personal and financial information after a person becomes a bank customer. After that, the bank can share the data with the recommendation

system. It will enter the next section – monetization. The system will generate analytical reports and provide a list of recommended products for each customer. Above outputs will be sent back to a virtual bank so that it can understand its customers' preference, implement new product/service (if necessary) and deliver personalized marketing. Based on the personalized recommendation, it is expected that the customer will more likely to be attracted by the advertisement and purchase the financial product/service. Virtual bank can generate more income under this circumstance. Subsequently, virtual bank will retrieve customer transaction data, implying the start of maintenance section. Recommendation system will do the same jobs as previous but the data contains not only customer personal and financial information but also transaction data. It can enrich the modeling and provide updated results to virtual bank. Bank can keep track of customer behavior and change the marketing if needed.

Apart from marketing to existing customer, virtual bank can utilize the data to advertise their financial product/service to potential target customers.

3.3. System Architecture

Recommendation system is developed by several steps (see ***Appendix 10.6***). After receiving customer data from virtual bank, the first step is data cleaning, including handling with missing data, dropping duplicated records and/or the fields no longer required. Then, it is feature engineering. Continuous variables are binned so that they can be easily scaled in the distance matrix later. Two similarity matrixes derived from memory-based collaborative filtering and demographic filtering are produced afterwards. Two matrixes will then be combined and build a recommendation system to output the probability each customer will purchase the products and determine the recommended product based on the probability result. The details will be explained in Section 5.

Section 4. Dataset

4.1. Overview

For demonstration purpose in project, customer behavior data is collected from open-source platform called Kaggle [8]. In reality, the customer data should be provided by the virtual bank so that the recommendation system can be built specifically for them.

The dataset contains more than 14 million records of customers behavior data from Santander bank over 1.5 years (I.e., Jan 2015 – Jun 2016, see ***Appendix 10.7***). It involves nearly 1 million customer and 24 financial products, for example, "credit card", "securities", "mortgage". These products are the columns named: ind_(<*product*>)_ult1 (see ***Appendix 10.8****)*. The sample data is split into two sets – training set and testing set according to the record date (see ***Appendix 10.7***)

4.2. Variables

There are 48 fields in the database (see ***Appendix 10.8***), 2 of which are identifier, 22 of them are demographic data and the rest of them are the financial products.

Section 5. Method

5.1. Data Cleaning

In the dataset, there are missing values in 19 fields (see ***Appendix 10.9.1***). The missing values can be classified into three types. Below is the handling of each missing variable type:

1) Factor Variable: It accounts for a small subset of data only so the most common factor level will be inputted. If this handling will imbalance factor class, a new factor level will be set.
2) Numerical Variable: Setting the missing value to the average for each province.
3) Product Variable: As some customers might not eligible to purchase certain financial products, there are "NA" in the database. Zero will be set to indicate no product ownership.

5.2. Feature Engineering

According to OmniSci [11], feature engineering refers to the preprocessing step to transform the raw data into the relevant variables that can be utilized in machine learning. It is important when creating a predictive model as it can improve the machine learning algorithm performance. In this project, the continuous variables, for example, “renta” (I.e., gross income), “age” and “antiguedad” (I.e., customer seniority), are binned so that all variables, especially demographic and ownership variables, can be treated as binary and easily scaled in distance matrix.

5.3. Designing Product Recommendation System

5.3.1. Data Subset

Since the dataset is quite large, only the months of interest is extracted for prediction. The records in three month (May 2015, June 2015 and May 2016) are used in the model to predict which financial products each customer will be likely to purchase in June 2016.

5.3.2. Similarity Matrix

Dataset is then split into training set and validation set with an 80:20 ratio. Two similarity matrixes are then calculated under memory-based collaborative filtering and demographic filtering. To avoid predicting the products already purchased by customer, the corresponding probabilities will be nullified. After that, weighted average prediction is derived using the probabilities from the above matrixes.

5.3.3. Optimizing Matrix Parameter

To select the optimal predictive model, the models with different weight ratio between memory-based collaborative filtering and demographic filtering are evaluated using validation set. Top 7 recommended products will be compared with the real transaction record to calculate the precision. The higher precision, the better model performance. According to ***Appendix 10.10***, when memory-base collaborative filtering and demographic filtering are combined with a ratio of 60:40, the model performed the best. Therefore, this weighting will be used as recommendation model in this project.

5.3.4. Output

Two outputs will be generated from the recommendation system and sent to virtual bank. First, it is the financial product purchase probability of each customer (see ***Appendix 10.2.1***). Second, it is the top N recommended product of each customer based on the above probability, where N depends on the virtual banks’ requirement (see ***Appendix 10.2.2***).

Section 6. Data Visualization

Data visualization is provided based on the customer data from the bank, including demographic data and the product purchase history. The data is first categorized based on customer’s characteristics and then analyzed to do summary and find the pattern if any. After that the information will be stored in two reports – customer overview report and customer detailed report (see Section 7 for details) according to the level of data granularity. Both reports give virtual bank a full picture of its customer’s status and predicted financial behavior. Virtual bank can know its customer better, and sell additional product to their customers effectively and precisely in the coming future.

Section 7. Analytical Reporting

7.1. Customer Overview Report

After receiving customer data, they will be stored in MySQL-based database. Analytical reports will be generated automatically using R Language and Python and sent to virtual bank for their further analysis. The graphs and reports can be found in ***Appendix 10.11.1***.

7.1.1. Report Generated by R Language

It shows a snapshot of original customer data. There are five sessions:

1.) Product Correlation
2.) Age
3.) Gender
4.) Gross Income
5.) Customer Type

Below is the analysis result for the dataset used in this project:

In part 1, it investigated the products customer purchased in the past and checked if there were association between the products. It is important for marketing product as if there is strongly relationship of two products (e.g., item A and item B), it means that when customer purchases item A, he/she is more likely to buy item B as well. From the correlation matrix, there is positively correlation between some products, such as direct debit, payroll and pension. Moreover, current account was negatively correlated to payroll account and pension.

In part 2, the customer's age distribution and the purchasing power of each age group is reviewed. Customers mainly aged between 20 and 40. Customers aged 41-50 purchased the financial products the most and those aged 21-30 have been purchased more products during the period of time.

In part 3, the customer's gender distribution in total and in each product is analyzed. The time-series purchasing power analysis of each gender is also performed. It is found that the number of products purchased by both female and male have been increasing. Female utilized more financial services than male in each product type.

In part 4, the relationship between average number of product and customer's gross income is investigated. In the dataset, most of the customers didn't have a huge gross income. However, the gross income is not a significant factor to determine the number of financial products to be purchased by the customer.

In part 5, the statistics of customer type is summarized. Most of the customers are individuals. During the 3rd quarter of 2015, the number of products utilized by graduate student was increased suddenly.

7.1.2. Report Generated by Python

In the report generated by Python, it mainly focuses on the result derived from the above product recommendation system. There are four areas:

1.) Product Usage Rate of Customer
2.) Top 3 Recommended Products for Each Customer Group
3.) Distribution of Recommended Products
4.) Distribution of Customer (Credit Card Product only)

Below is the analysis result for the dataset used in this project:

In part 1, it showed the product usage rate of customer. Customer utilized current account mostly during the period of time

In part 2, it listed out the top 3 recommended product for each customer group (I,e, VIP, individuals and graduate student). For customer which is classified as the group of individuals or graduate students, 1st recommended product is pensions. For the VIP client, e-account is highly recommended based on the recommendation system.

In part 3, it summarized the distribution of recommended product among all customers. Pensions was mostly recommended to the customers.

In part 4, the distribution of customer with a recommended product of credit card is shown. Most of them are male and aged between 30-40.

7.2. Customer Detailed Report

This report is tailored to each customer individually. Through this customer report, the virtual bank can quickly learn the potential marketing direction of each customer, which facilitates the marketing staff to target marketing to get customers to invest more assets into the bank to increase customer satisfaction with the Virtual Bank products. The report is divided into the following three sections (Sample report can be found in ***Appendix 10.11.2)***.

1.) Customer Information
2.) Training Result
3.) Recommended Product

From the first part, it lists out the demographic data of individual customers, such as age, gender, and annual household income. Virtual bank can then learn immediately whether the customer is an employee of the company and whether he or she is a new customer within 6 months, etc.

The second part of the report shows the result of our analysis and recommendation system. It consists of two fields: customer account histories data (that is bank accounts that the user has opened with the bank in the past, refer to the field “Accounts in Records” in ***Appendix 10.11.2***) and the recommended product type (refer to the field “Marketing Direction” in ***Appendix 10.11.2***.)

The last part of the report facilitates the bank's marketing staff to contact the customer directly with information about the product being promoted. As we assume that the system is built for ZA Bank for illustration, the recommendations in the sample report are all ZA Bank products. In case other virtual banks are interested in accessing our recommendation system, it can be easily deployed using the APIs already established for this project.

For ZA Bank, it is worth noting that the product portals for Zhongan Bank are only ZA Bank and ZA One, so we will target different marketing directions, for example, the algorithm in the previous section calculates those different customers have long-term and short-term deposit preferences, so the product recommendation section will be differentiated. The product recommendations are differentiated in the product recommendation section, prompting the virtual bank's marketers to steer customers towards short-term (3-month, 6-month) fund products with higher yields or long-term (3-year, 5-year) fund products with higher yields.

Section 8. Conclusion

8.1. Summary

Virtual banks are constantly attacking the strongest players in the traditional industry, creating one new business and product after another in the wave of the digital economy and the booming development of financial technology.

Virtual bank can bring positive outcome to different parties. From the perspective of macro-economy, the gradual formation of healthy competition has provided greater market potential; for investors, the innovative ability of each virtual bank has greatly enhanced shareholders' confidence in investment; for consumers, the continuous innovation of products and services by virtual bank practitioners has accelerated the process of "de-banking" and made intelligent financial services within reach. However, the market penetration rate of virtual banking is still low in Hong Kong.

In order to solve the issue, our project aims to investigate how to accelerate customer acquisition and retention in digital banking by improving customer experience using AI and big data. We built a personalized recommendation system and delivered two customer related analytical reports for virtual banks. It is expected to give them a better idea of what their clients want and direction to improve. It will then help virtual banks to deliver a highly customized cyber experience for their clients, differentiate themselves from traditional banks and drive rapid market penetration in Hong Kong.

8.2. Limitations

8.2.1. Out-dated Customer Data

The dataset used in this study is scrapped from website and the customer purchase records are not the latest snapshot (I.e., Jun 2016). Since the customer purchase record changes over the time. the modeling required to be completed with the latest dataset in a timely manner to perform prediction precisely.

8.2.2. Non-time Weighted Modeling

In the modelling stage, a hybrid filtering approach is used and the records in different period are assumed to affect customer's preference equally. In other words, the record was not weighted according to the corresponding date during modeling. In fact, the more recent purchase record can better predict the customer's preference.

8.2.3. Modeling without Risk Appetite Consideration

The client's appetite for risk is not considered in the modeling, so the recommended products based on the historical record, may not be fully suitable for that client.

8.2.4. Partially Automated Reporting

The overview analytical reports were automated by R language and Python while the customer detailed reports were manually generation due to the report complexity.

8.3. Future Work

8.3.1. Real Time Capability

As mentioned above (Section 8.2.1), the prediction requires to be time-critical so that the virtual bank can gain prompt insight, update its marketing based on the change in customer's preference and make the business decision quickly. It is suggested to product the model to re-forecast in real time and fine-tine the model continuously to improve the model accuracy.

8.3.2. Cross-Validation

The dataset is split into training set and validation set with a ratio of 80:20 to build the predictive model. In fact, the raw data can be repeatedly split into the two subsets. This process is called cross-validation, which can utilize all significant data for training purpose and retrieve better and more stable result.

8.3.3. Time-Weighted Modeling

The more recent observation can be assigned to have more weight in the modeling. It can also consider the seasonality and holiday effect on customer's purchasing

8.3.4. Scope of Financial Product

It will be more practical to analyze competitive differentiation in the context of investment and market preferences and create some innovations in terms of virtual banking instead of optimizing and/or recommending the existing products.

Section 9. Reference

[1] Hong Kong Monetary Authority, “Hong Kong Monetary Authority - Annual Report 2020”, Hong Kong, 2020, https://www.hkma.gov.hk/eng/data-publications-and-research/publications/annual-report/2020/

[2] Business Jargons, “What Is Virtual Banking? Definition and Meaning”, 14 Jan 2016, https://businessjargons.com/virtual-banking.html [Accessed 6 Oct 2021]

[3] KPMG, “Embracing change, driving growth: Hong Kong Banking Report 2021”, Jun 2021, https://assets.kpmg/content/dam/kpmg/cn/pdf/en/2021/06/hong-kong-banking-report-2021.pdf

[4] Vpon, “The New Wave of Tech Behemoths in Hong Kong”, Jun 2021, https://www.vpon.com/wp-content/uploads/The_New_Wave_of_Tech_Behemoths_in_Hong_Kong_EN.pdf

[5] Benjamin Quinlan & Eashan Trehan, “Branching Off - The Outlook For Hong Kong’s Virtual Banks”, Mar 2021, https://www.quinlanandassociates.com/wp-content/uploads/2021/03/Quinlan-Associates-Branching-Off.pdf

[6] Iateilang Ryngksai & L. Chameikho, “Recommender Systems: Types of Filtering Techniques”, Nov 2014, https://www.ijert.org/research/recommender-systems-types-of-filtering-techniques-IJERTV3IS110197.pdf

[7] Richard Laycock & Susannah Binsted, “Digital Only Banking Adoption 2021”, Oct 2021, https://www.finder.com/hk/virtual-banking-statistics

[8] Banco Santander, “Santander Product Recommendation Can you pair products with people?”, Jun 2016, https://www.kaggle.com/c/santander-product-recommendation/

[9] P. H. Aditya, I. Budi, Q. Munajat, “A Comparative Analysis of Memory-based and Model-based Collaborative Filtering on the Implementation of Recommender System for Ecommerce in Indonesia : A Case Study PT X”, Nov 2014, https://goribmunajat.github.io/files/comparative-analysis-memory-based-model-based-recommendation-systems.pdf

[10] Tarang Shah, “About Train, Validation and Test Sets in Machine Learning”, Dec 2017, https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

[11] OmniSci, “Feature Engineering”, 2021, https://www.omnisci.com/technical-glossary/feature-engineering

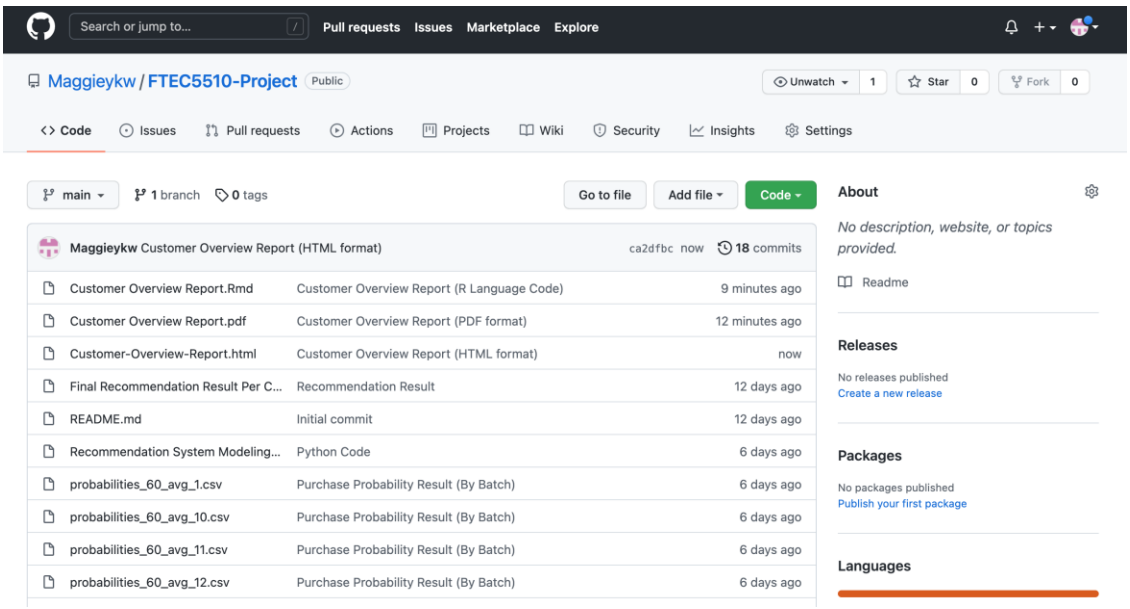
[12] HKMA, ‘Monthly Statistical Bulletin’, 2021, https://www.hkma.gov.hk/eng/data-publications-andresearch/data-and-statistics/monthly-statistical-bulletin/

[13] HKMA, ‘List of Licensed Banks’, November 2020, https://www.hkma.gov.hk/media/eng/doc/key-functions/bankingstability/banking-policy-and-supervision/list_of_lb.xls

Section 10. Appendix

10.1. Code for Recommendation System Modeling

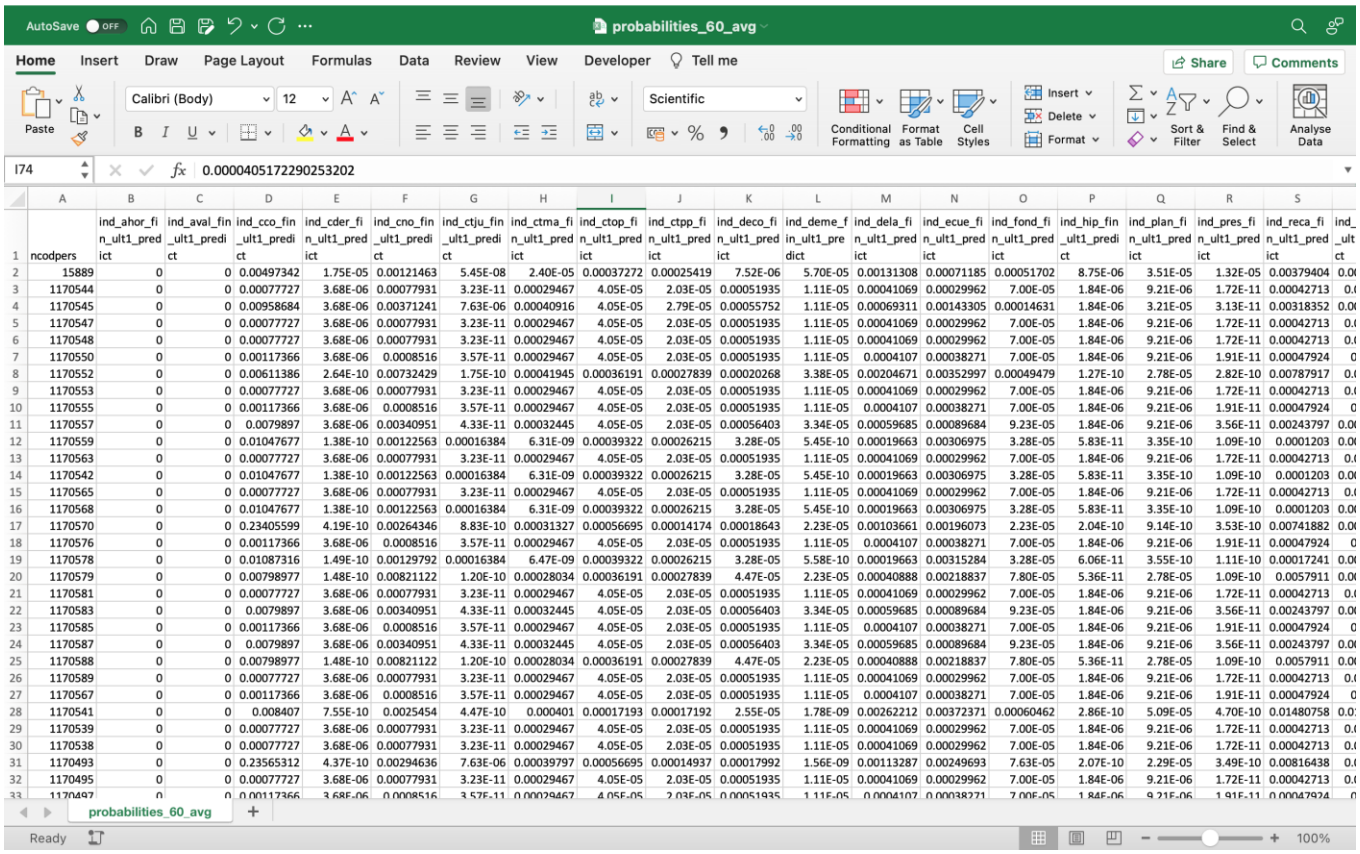
The code is uploaded to Github (https://github.com/Maggieykw/FTEC5510-Project.git) and the file name is "Recommendation System Modeling.ipynb". A capture of the Github files is shown below.



10.2. Output from Recommendation System

10.2.1. Purchase Probabilities of Each Customer

The probability result is uploaded to Github by batch (https://github.com/Maggieykw/FTEC5510-Project.git) and the file name is "probabilities_60_avg_<*batch no*>.csv". A capture of the file is shown below.



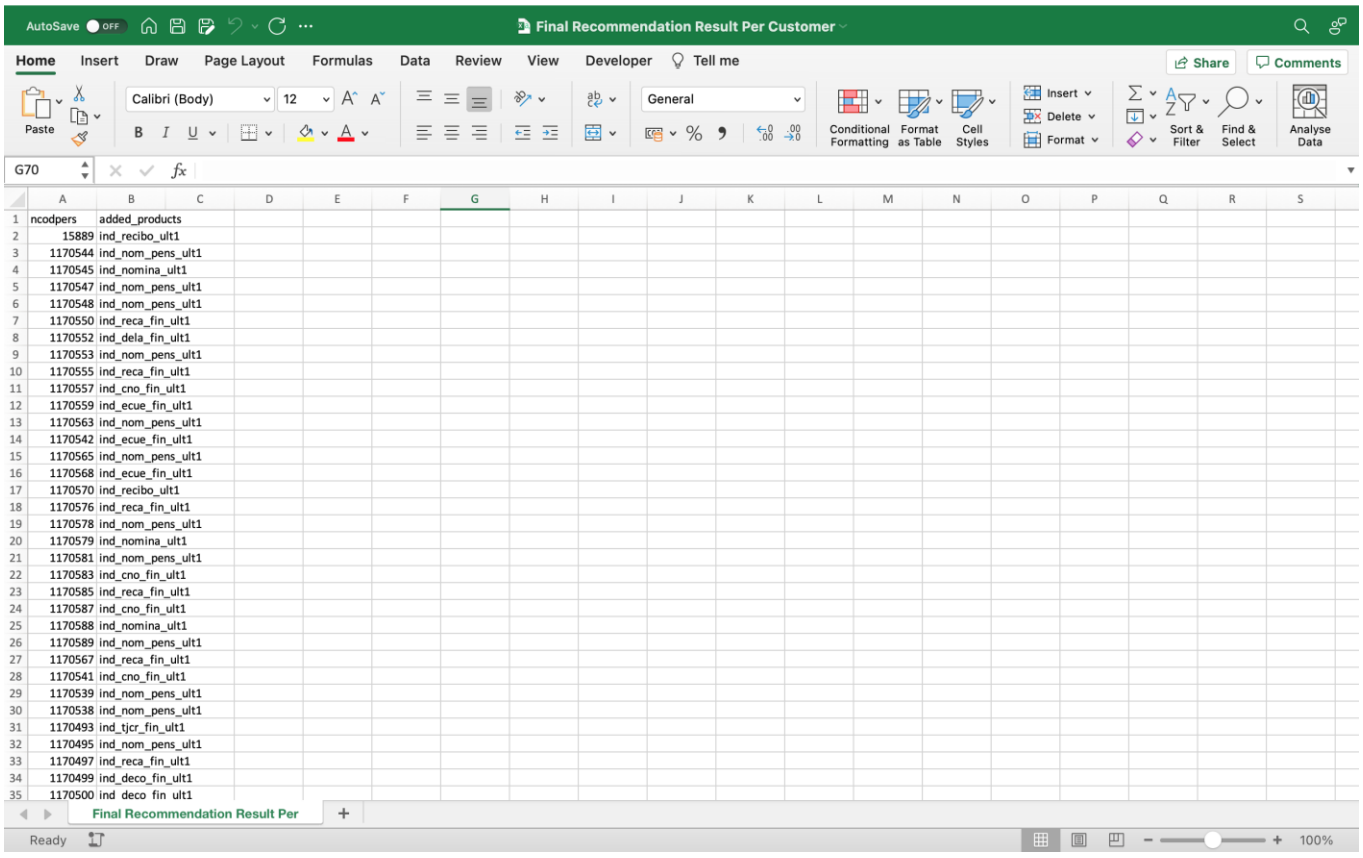
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	ncodpers	ind_ahor_fin_ult1_predict	ind_aval_fin_ult1_predict	ind_cco_fin_ult1_predict	ind_cder_fin_ult1_predict	ind_cno_fin_ult1_predict	ind_ctju_fin_ult1_predict	ind_ctma_fin_ult1_predict	ind_ctop_fin_ult1_predict	ind_ctpp_fin_ult1_predict	ind_deco_fin_ult1_predict	ind_deme_fin_ult1_predict	ind_dela_fin_ult1_predict	ind_ecue_fin_ult1_predict	ind_fond_fin_ult1_predict	ind_hip_fin_ult1_predict	ind_plan_fin_ult1_predict	ind_pres_fin_ult1_predict	ind_reca_fin_ult1_predict	ind_ult
2	15889	0	0	0.00497342	1.75E-05	0.00121463	5.45E-08	2.40E-05	0.00037272	0.00025419	7.52E-06	5.70E-05	0.00131308	0.00071185	0.00051702	8.75E-06	3.51E-05	1.32E-05	0.00379404	0.0
3	1170544	0	0	0.00077727	3.68E-06	0.00077931	3.23E-11	0.00029467	4.05E-05	2.03E-05	0.00051935	1.11E-05	0.00041069	0.00029962	7.00E-05	1.84E-06	9.21E-06	1.72E-11	0.00042713	0.0
4	1170545	0	0	0.00958684	3.68E-06	0.00371241	7.63E-06	0.00040916	4.05E-05	2.79E-05	0.00055752	1.11E-05	0.00069311	0.00143305	0.00014631	1.84E-06	3.21E-05	3.13E-11	0.00318352	0.0
5	1170547	0	0	0.00077727	3.68E-06	0.00077931	3.23E-11	0.00029467	4.05E-05	2.03E-05	0.00051935	1.11E-05	0.00041069	0.00029962	7.00E-05	1.84E-06	9.21E-06	1.72E-11	0.00042713	0.0
6	1170548	0	0	0.00077727	3.68E-06	0.00077931	3.23E-11	0.00029467	4.05E-05	2.03E-05	0.00051935	1.11E-05	0.00041069	0.00029962	7.00E-05	1.84E-06	9.21E-06	1.72E-11	0.00042713	0.0
7	1170550	0	0	0.00117366	3.68E-06	0.0008516	3.57E-11	0.00029467	4.05E-05	2.03E-05	0.00051935	1.11E-05	0.0004107	0.00038271	7.00E-05	1.84E-06	9.21E-06	1.91E-11	0.00047924	0
8	1170552	0	0	0.00611386	2.64E-10	0.00732429	1.75E-10	0.00041945	0.00036191	0.00027839	0.00020268	3.38E-05	0.00204671	0.00352997	0.00049479	1.27E-10	2.78E-05	2.82E-10	0.00787917	0.0
9	1170553	0	0	0.00077727	3.68E-06	0.00077931	3.23E-11	0.00029467	4.05E-05	2.03E-05	0.00051935	1.11E-05	0.00041069	0.00029962	7.00E-05	1.84E-06	9.21E-06	1.72E-11	0.00042713	0.0
10	1170555	0	0	0.00117366	3.68E-06	0.0008516	3.57E-11	0.00029467	4.05E-05	2.03E-05	0.00051935	1.11E-05	0.0004107	0.00038271	7.00E-05	1.84E-06	9.21E-06	1.91E-11	0.00047924	0
11	1170557	0	0	0.0079897	3.68E-06	0.00340951	4.33E-11	0.00032445	4.05E-05	2.03E-05	0.00056403	3.34E-05	0.00059685	0.00089684	9.23E-05	1.84E-06	9.21E-06	3.56E-11	0.00243797	0.0
12	1170559	0	0	0.01047677	1.38E-10	0.00122563	0.00016384	6.31E-09	0.00039322	0.00026215	3.28E-05	5.45E-10	0.00019663	0.00306975	3.28E-05	5.83E-11	3.35E-10	1.09E-10	0.0001203	0.0
13	1170563	0	0	0.00077727	3.68E-06	0.00077931	3.23E-11	0.00029467	4.05E-05	2.03E-05	0.00051935	1.11E-05	0.00041069	0.00029962	7.00E-05	1.84E-06	9.21E-06	1.72E-11	0.00042713	0.0
14	1170542	0	0	0.01047677	1.38E-10	0.00122563	0.00016384	6.31E-09	0.00039322	0.00026215	3.28E-05	5.45E-10	0.00019663	0.00306975	3.28E-05	5.83E-11	3.35E-10	1.09E-10	0.0001203	0.0
15	1170565	0	0	0.00077727	3.68E-06	0.00077931	3.23E-11	0.00029467	4.05E-05	2.03E-05	0.00051935	1.11E-05	0.00041069	0.00029962	7.00E-05	1.84E-06	9.21E-06	1.72E-11	0.00042713	0.0
16	1170568	0	0	0.01047677	1.38E-10	0.00122563	0.00016384	6.31E-09	0.00039322	0.00026215	3.28E-05	5.45E-10	0.00019663	0.00306975	3.28E-05	5.83E-11	3.35E-10	1.09E-10	0.0001203	0.0
17	1170570	0	0	0.23405599	4.19E-10	0.00264346	8.83E-10	0.00031327	0.00056695	0.00014174	0.00018643	2.23E-05	0.00103661	0.00196073	2.23E-05	2.04E-10	9.14E-10	3.53E-10	0.00741882	0.0
18	1170576	0	0	0.00117366	3.68E-06	0.0008516	3.57E-11	0.00029467	4.05E-05	2.03E-05	0.00051935	1.11E-05	0.0004107	0.00038271	7.00E-05	1.84E-06	9.21E-06	1.91E-11	0.00047924	0
19	1170578	0	0	0.01087316	1.49E-10	0.00129792	0.00016384	6.47E-09	0.00039322	0.00026215	3.28E-05	5.58E-10	0.00019663	0.00315284	3.28E-05	6.06E-11	3.55E-10	1.11E-10	0.00017241	0.0
20	1170579	0	0	0.00798977	1.48E-10	0.00821122	1.20E-10	0.00028034	0.00036191	0.00027839	4.47E-05	2.23E-05	0.00040888	0.00218837	7.80E-05	5.36E-11	2.78E-05	1.09E-10	0.0057911	0.0
21	1170581	0	0	0.00077727	3.68E-06	0.00077931	3.23E-11	0.00029467	4.05E-05	2.03E-05	0.00051935	1.11E-05	0.00041069	0.00029962	7.00E-05	1.84E-06	9.21E-06	1.72E-11	0.00042713	0.0
22	1170583	0	0	0.0079897	3.68E-06	0.00340951	4.33E-11	0.00032445	4.05E-05	2.03E-05	0.00056403	3.34E-05	0.00059685	0.00089684	9.23E-05	1.84E-06	9.21E-06	3.56E-11	0.00243797	0.0
23	1170585	0	0	0.00117366	3.68E-06	0.0008516	3.57E-11	0.00029467	4.05E-05	2.03E-05	0.00051935	1.11E-05	0.0004107	0.00038271	7.00E-05	1.84E-06	9.21E-06	1.91E-11	0.00047924	0
24	1170587	0	0	0.0079897	3.68E-06	0.00340951	4.33E-11	0.00032445	4.05E-05	2.03E-05	0.00056403	3.34E-05	0.00059685	0.00089684	9.23E-05	1.84E-06	9.21E-06	3.56E-11	0.00243797	0.0
25	1170588	0	0	0.00798977	1.48E-10	0.00821122	1.20E-10	0.00028034	0.00036191	0.00027839	4.47E-05	2.23E-05	0.00040888	0.00218837	7.80E-05	5.36E-11	2.78E-05	1.09E-10	0.0057911	0.0
26	1170589	0	0	0.00077727	3.68E-06	0.00077931	3.23E-11	0.00029467	4.05E-05	2.03E-05	0.00051935	1.11E-05	0.00041069	0.00029962	7.00E-05	1.84E-06	9.21E-06	1.72E-11	0.00042713	0.0
27	1170567	0	0	0.00117366	3.68E-06	0.0008516	3.57E-11	0.00029467	4.05E-05	2.03E-05	0.00051935	1.11E-05	0.0004107	0.00038271	7.00E-05	1.84E-06	9.21E-06	1.91E-11	0.00047924	0
28	1170541	0	0	0.008407	7.55E-10	0.0025454	4.47E-10	0.000401	0.00017193	0.00017192	2.55E-05	1.78E-09	0.00262212	0.00372371	0.00060462	2.86E-10	5.09E-05	4.70E-10	0.01480758	0.0
29	1170539	0	0	0.00077727	3.68E-06	0.00077931	3.23E-11	0.00029467	4.05E-05	2.03E-05	0.00051935	1.11E-05	0.00041069	0.00029962	7.00E-05	1.84E-06	9.21E-06	1.72E-11	0.00042713	0.0
30	1170538	0	0	0.00077727	3.68E-06	0.00077931	3.23E-11	0.00029467	4.05E-05	2.03E-05	0.00051935	1.11E-05	0.00041069	0.00029962	7.00E-05	1.84E-06	9.21E-06	1.72E-11	0.00042713	0.0
31	1170493	0	0	0.23565312	4.37E-10	0.00294636	7.63E-06	0.00039797	0.00056695	0.00014937	0.00017992	1.56E-09	0.00113287	0.00249693	7.63E-05	2.07E-10	2.29E-05	3.49E-10	0.00816438	0.0
32	1170495	0	0	0.00077727	3.68E-06	0.00077931	3.23E-11	0.00029467	4.05E-05	2.03E-05	0.00051935	1.11E-05	0.00041069	0.00029962	7.00E-05	1.84E-06	9.21E-06	1.72E-11	0.00042713	0.0
33	1170497	0	0	0.00117366	3.68E-06	0.0008516	3.57E-11	0.00029467	4.05E-05	2.03E-05	0.00051935	1.11E-05	0.0004107	0.00038271	7.00E-05	1.84E-06	9.21E-06	1.91E-11	0.00047924	0

Customer ID	Short-term deposits	Long-term deposits	...	Mortgage	Credit Card	Securities
23890	0.00002229	0.00136929	...	0.00000926	0.00833274	0.55961979
67047	0.00002225	0.00137221	...	0.55790591	0.00835016	0.00033526
...	...	...	...	...	...	...
202697	0.00002411	0.56163096	...	0.00000831	0.00688280	0.00012965
314445	0.03379550	0.01150405	...	0.00000185	0.00055939	0.00005915
360164	0.00001717	0.00123101	...	0.00000805	0.58521574	0.00012405

(Subset of combined demographic-based and memory-based probabilities)

10.2.2. Top 1 Recommended Item for Each Customer

The probability result is uploaded to Github (https://github.com/Maggieykw/FTEC5510-Project.git) and the file name is “Final Recommendation Result Per Customer.csv”. A capture of the file is shown below.



ncodpers	added_products
15889	ind_recibo_ult1
1170544	ind_nom_pens_ult1
1170545	ind_nomina_ult1
1170547	ind_nom_pens_ult1
1170548	ind_nom_pens_ult1
1170550	ind_reca_fin_ult1
1170552	ind_dela_fin_ult1
1170553	ind_nom_pens_ult1
1170555	ind_reca_fin_ult1
1170557	ind_cno_fin_ult1
1170559	ind_ecue_fin_ult1
1170563	ind_nom_pens_ult1
1170542	ind_ecue_fin_ult1
1170565	ind_nom_pens_ult1
1170568	ind_ecue_fin_ult1
1170570	ind_recibo_ult1
1170576	ind_reca_fin_ult1
1170578	ind_nom_pens_ult1
1170579	ind_nomina_ult1
1170581	ind_nom_pens_ult1
1170583	ind_cno_fin_ult1
1170585	ind_reca_fin_ult1
1170587	ind_cno_fin_ult1
1170588	ind_nomina_ult1
1170589	ind_nom_pens_ult1
1170567	ind_reca_fin_ult1
1170541	ind_cno_fin_ult1
1170539	ind_nom_pens_ult1
1170538	ind_nom_pens_ult1
1170493	ind_tjcr_fin_ult1
1170495	ind_nom_pens_ult1
1170497	ind_reca_fin_ult1
1170499	ind_deco_fin_ult1
1170500	ind_deco_fin_ult1

10.3. Work Allocation and Self-Reflection

10.3.1. LIU Hao Fei (Business Analyst)

During this semester, I was lucky enough to work with three other group members to complete a project on financial infrastructure. With four members from different universities and different academic backgrounds, we worked together to create a product recommendation system for virtual banking. Through my communication with the group members, my course supervisor Professor Chan and the developers of the Simnectz platform, I was able to learn not only about Fintech in my field, but also to gain a better understanding of the latest developments in the financial industry in Hong Kong. And because we chose a

virtual banking product, this was an opportunity for me to gain a deeper understanding of the local Fintech development in Hong Kong.

In the past two months, all four of us contributed a lot, from the initial research on the Hong Kong FinTech market, the establishment of the project proposition and the development of the entire product. For me personally, it was rewarding to strengthen my proficiency in MySQL and the Python language. Through the group activities, I learnt about the construction of products within the bank and the techniques used to provide services to the providers of financial services.
Overall, this project has strengthened my understanding of the financial and technical fields and my proficiency in programming languages, which is exactly what I had hoped for before entering my Master's degree in FinTech at CUHK.

10.3.2. RAO Jia Yi (Researcher)

In this term, I have a chance to cooperate with my teammates to do the FTEC5510 project. By observing several types of Fintech topics in our lecture, our group was interested in the virtual banking industry and started to investigate in terms of this topic.

First, I did some research about the information of virtual banks in Hong Kong through HKMA and some financial report. Then I shared a brief overview of the history and development of virtual banking in Hong Kong with my teammates. We found that virtual banks all failed to maintain their clients and their services were quite limited. Only by putting more effort on customer acquisition and retention can the virtual banks be more defensible among the banking industry. Therefore, we found this issue interesting and meaningful and decided to build the recommendation system and generate data report to virtual banks. During the process of modelling, although I was not good at coding, I made some suggestions, asked my teammates for technical problems on python and sought to improve myself.

Finally, I have read all references and have a deeper understanding of our group’s projects. Throughout the progress of the project, our team members have a clear division of labor, actively cooperate, and communicate in time to make the project progress in an orderly manner. This experience left a deep impression on me and made me progress. It will also help me study, work and live in the future.

10.3.3. YIM Ka Wai (Researcher, Developer, Business Analyst)

During this term, my teammates and I have had a change to participate in a group project. There had a diverse set of topics related to FinTech, our group decided to build a financial product recommendation system and do data visualization based on customer data, including personal information, financial details, and transaction record. I acted as a project organizer, researcher, developer, and business analyst in the early stage, I did some research about FinTech to review the pain point of virtual banks are facing and figure out an innovative idea that helps them to resolve the above challenge. Then I presented my idea with my teammates. After that, I helped to build a predictive model using Python to derive purchase probability for each customer. Finally, I did data analysis and visualization using R Language.

I enjoyed finding pattern in datasets and performing prediction using big data approach throughout the project. I am interested in data science and have been learning learned different kind of programming language, for example, Python R, SAS, C++, JavaScript.

One of the challenges is modeling. Previously, I also did recommendation system, for example, using historical quantitative data to predict the movie rating of user or video game sales in the future using logistic regression method. This time the modeling is much more complicated as there are qualitative data apart from quantitative data and I needed to consolidate both data into the model. It took time for me to perform feature engineering and decide the weight between these two data types using trial and error approach.

Apart from technical part, it is not easy to get all teammates are on the same page. Sometimes, there were misunderstanding among our team due to the difference in our academic background. It requires additional efforts to illustrate the entire system architecture and the techniques involved for those with a non-technical background to make sure them understand the full picture and know what task they will focus on.

We had a great and interesting project topic and eventually generated a personalized report for virtual bank based on the customer data and the recommendation system we developed. Due to limited time, this project is only an introduction to recommendation system for virtual bank. There are several future works can be done to improve the model performance and outcomes. For example, cross-validation, real-time data scraping, using alternative data (for example, social media), selecting certain variables into modeling, testing different distance metrics and demographic weighting, testing different predictive models, automating the report generation, performing time series analysis taking account of seasonality and holiday effect. Moreover, there is room to improve the intercultural and interpersonal communication skills to set a clean goal, plan better and run the project smoothly.

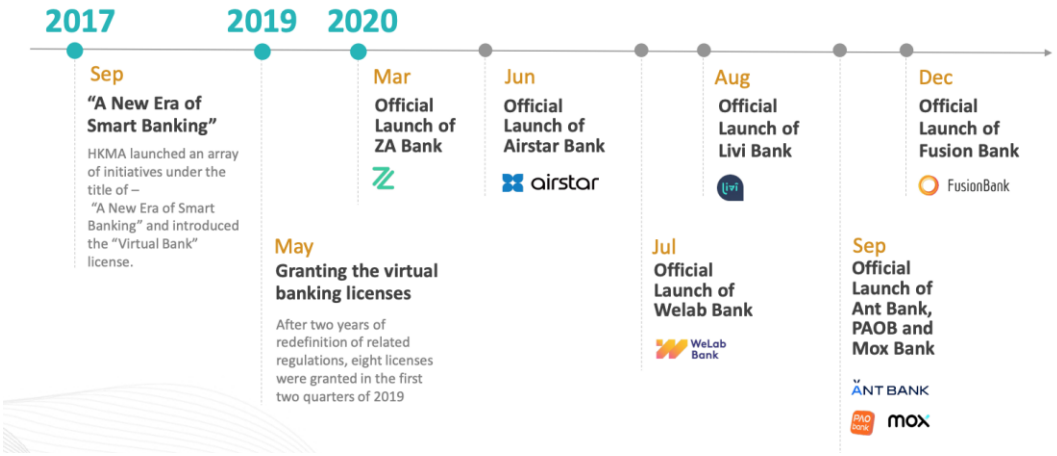
10.3.4. YU Tian Yi (Business Analyst)

Through group cooperation, I have gained a lot from this course. We cultivated our self-learning ability, cooperation consciousness and deepened our understanding of fintech projects. I became familiar with the entire process of agile project management and learned the importance of communication through group cooperation. Besides, I gained a better understanding of fintech. It is a solution that integrates technology, customer insight, financial scenarios and product operations to help financial institutions adapt to the new changes.

During the project, I participated in the background part with my teammates. We were interested in virtual banks in Hong Kong and decided to research the recommendation system of financial products. This field is inexperienced for me. I read the materials related to the logical structure and the architecture of the recommendation system, analysis of user behaviors, the calculation of similarity formulas between user items, etc. I explored the status of virtual banks and their financial products with my teammates. I tried exploratory data analysis which includes data screening, visualizing the data using histograms, boxplots and so forth to explore the relationships hidden behind the dataset. Data cleaning tasks which include handling the outliers and missing values were also conducted. Besides, I tried different BI tools (Tableau, Power BI) and Python visualization toolkits (Matplotlib, Seaborn, Plotly, Pyecharts), gaining a better understanding of how to design a good infographic using elements like type and color. To conclude, I harvested quite a lot through conducting this fintech project. It is beneficial for pursuing a career in data analysis and business intelligence in the future. Also, I feel very lucky to know and cooperate with my team members.

10.4. Figures

10.4.1. Current Development of Virtual Banks in Hong Kong



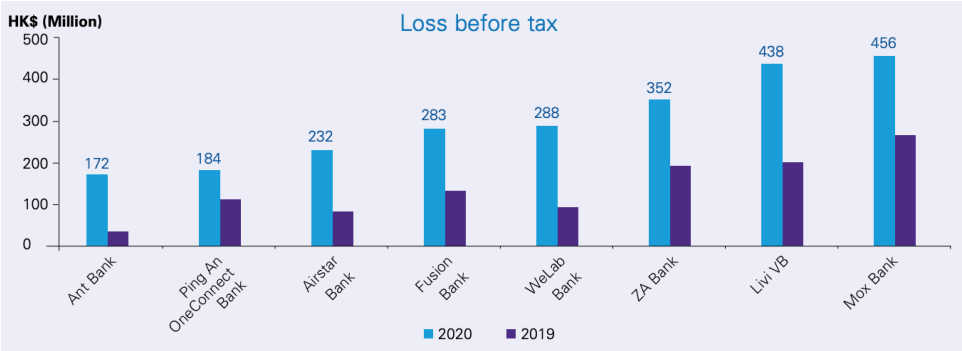
(Source: Vpon Virtual Banking Insight Report, 2021 [4])

10.4.2. List of Virtual Banks in Hong Kong

Virtual bank	Launch date
ZA Bank	March 2020
Airstar Bank	June 2020
WeLab Bank	July 2020
Livi VB	August 2020
Mox Bank	September 2020
Ant Bank	September 2020
Ping An OneConnect Bank	September 2020
Fusion Bank	December 2020

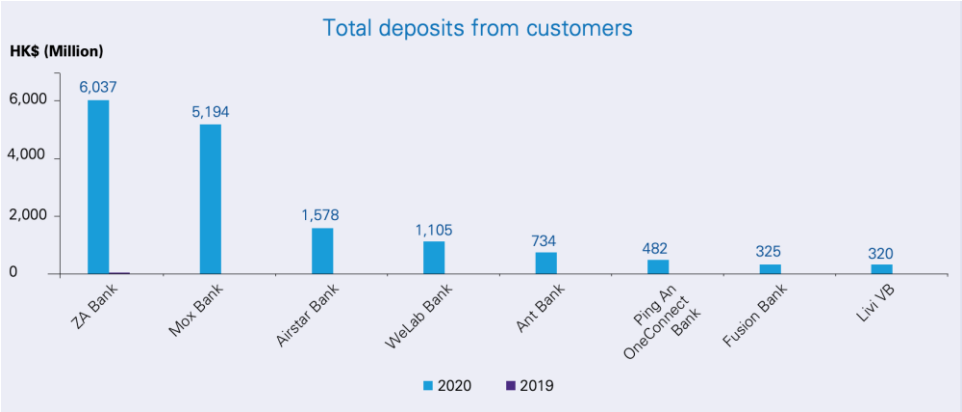
(Source: KPMG Hong Kong Banking Report, 2021 [3])

10.4.3. Loss Before Tax of Virtual Bank in Hong Kong



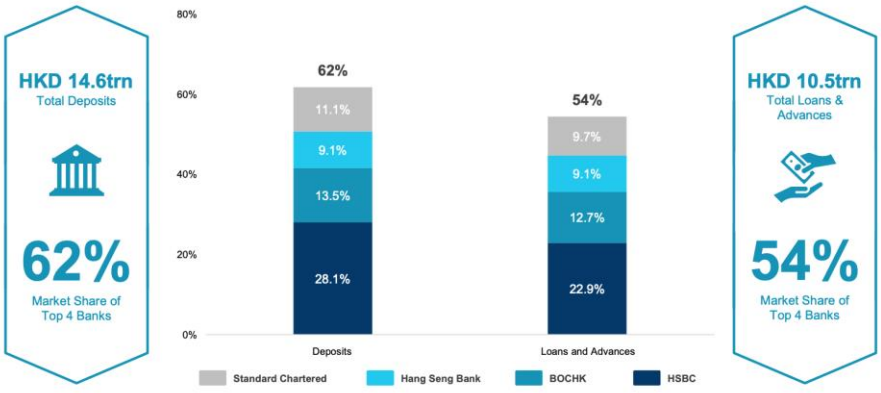
(Source: KPMG Hong Kong Banking Report, 2021 [3])

10.4.4. Total Customer Deposits of Virtual Bank in Hong Kong



(Source: KPMG Hong Kong Banking Report, 2021 [3])

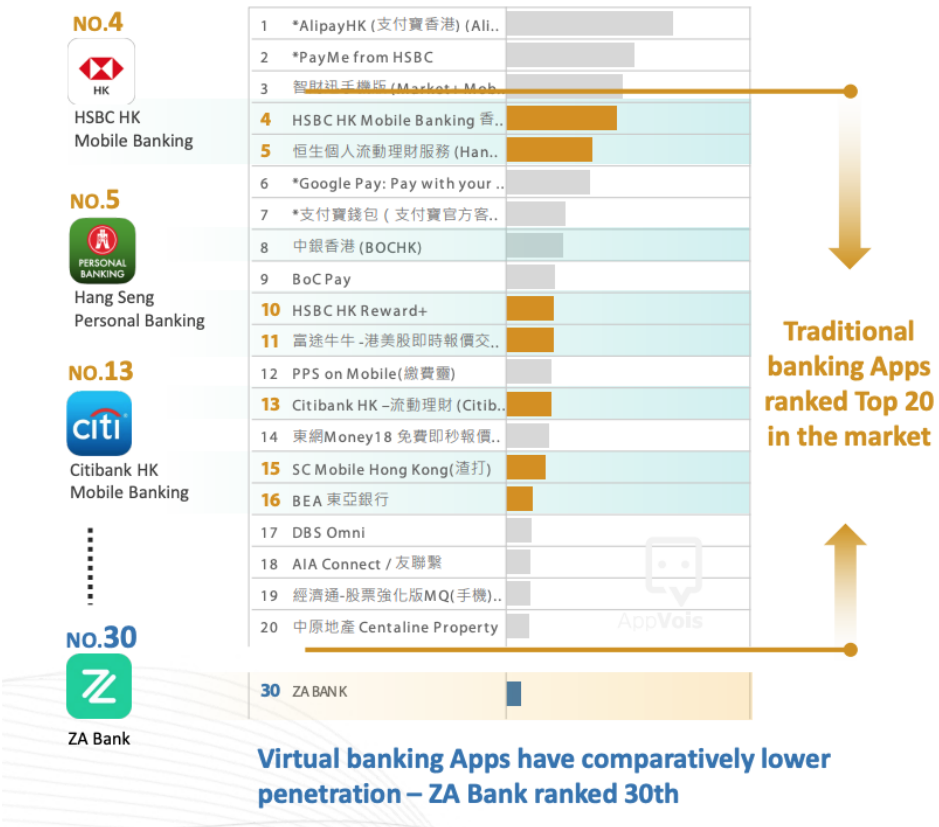
10.4.5. Hong Kong Banking Industry Market Share



	Deposits	Loans and Advances	Non-Interest Income*	Revenue*
HSBC	HKD 4.1 trillion	HKD 2.4 trillion	HKD 40.8 billion (42.8%)	HKD 95.35 billion
BOCHK	HKD 2.0 trillion	HKD 1.3 trillion	HKD 10.3 billion (28.0%)	HKD 36.80 billion
Stan Chart	HKD 1.6 trillion	HKD 1.0 trillion	HKD 20.7 billion (44.7%)	HKD 46.23 billion
Hang Seng	HKD 1.3 trillion	HKD 1.0 trillion	HKD 25.2 billion (52.2%)	HKD 48.22 billion

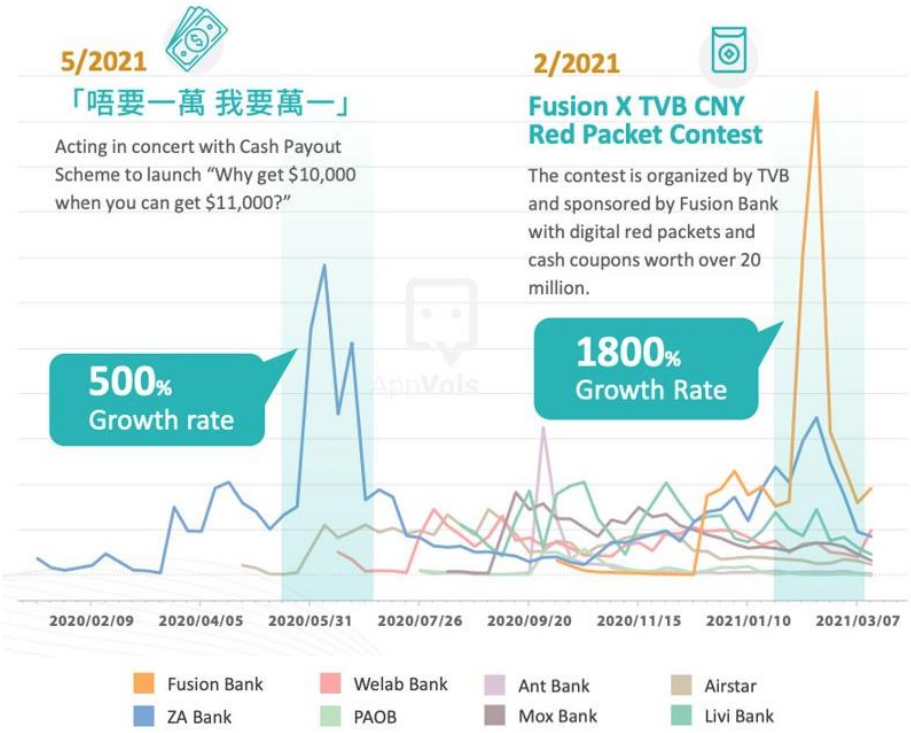
(Source: Quinlan&Associates Report, 2021 [5])

10.4.6. Overview of Hong Kong Banking and Finance App Industry



(Source: Vpon Virtual Banking Insight Report, 2021 [4])

10.4.7. Virtual Banking Apps New Install Trend in Hong Kong



(Source: Vpon Virtual Banking Insight Report, 2021 [4])

10.4.8. Services Provided by Virtual Banks in Hong Kong

Virtual Bank	ZA Bank	Airstar Bank	WeLab Bank	livi Bank	Mox Bank	Ant Bank	PAOB	Fusion Bank
Launch Date	Mar 2020	Jun 2020	Jul 2020	Aug 2020	Sep 2020	Sep 2020	Sep 2020	Dec 2020
Deposits* (HKD '000)	2,757,955	356,578	187,289	183	28,030	1,334	6,591	N/A
Services								
Savings	✓	✓	✓	✓	✓	✓	✓	✓
Time Deposits	✓	✓	✓	×	×	×	×	✓
Personal Loans	✓	✓	×	×	×	×	×	×
Business Loans	✓	×	×	×	×	×	✓	×
Transfer / Pay	✓	✓	✓	✓	✓	✓	✓	✓
Debit Card	✓	×	✓	✓	✓	×	×	×
FX	×	×	×	×	×	×	×	✓
Insurance	✓	×	×	×	×	×	×	×
Budgeting Tools	×	×	×	×	✓	×	×	×

*As of June 2020

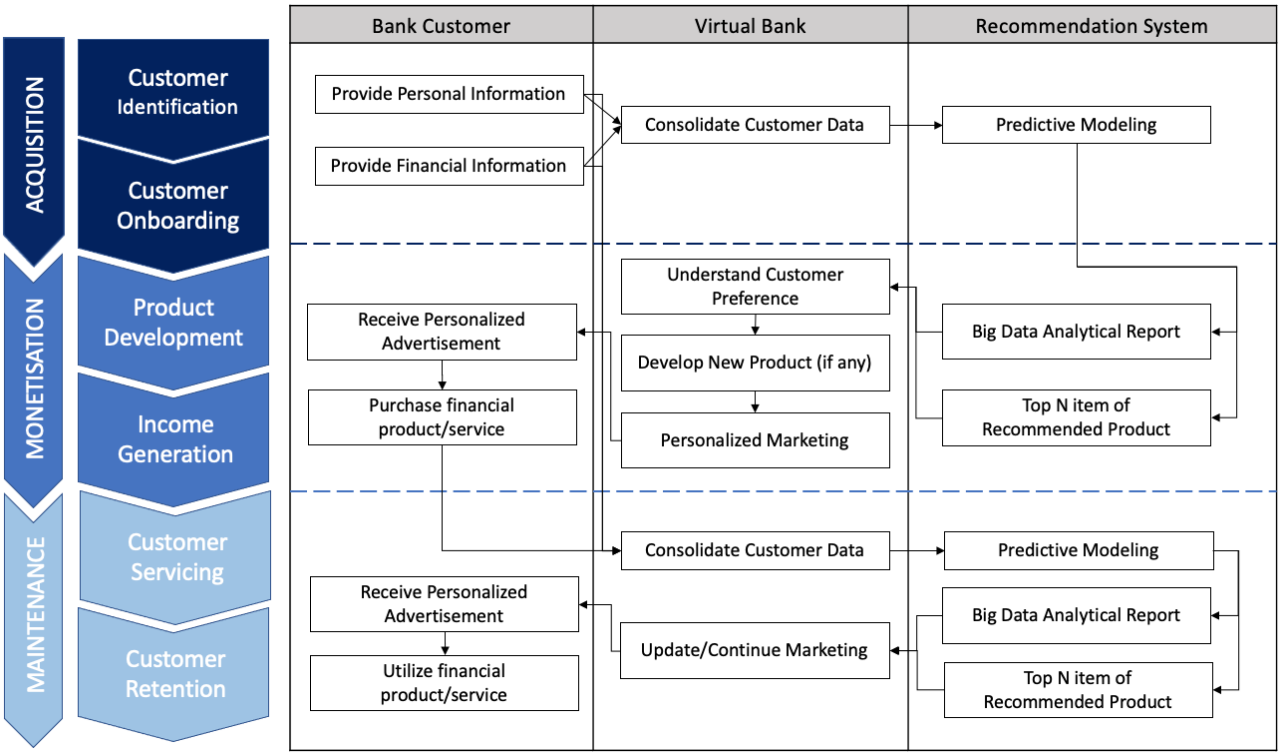
✓ Offered × Not Offered

(Source: Quinlan & Assoicates Analysis Report, 2021 [5])

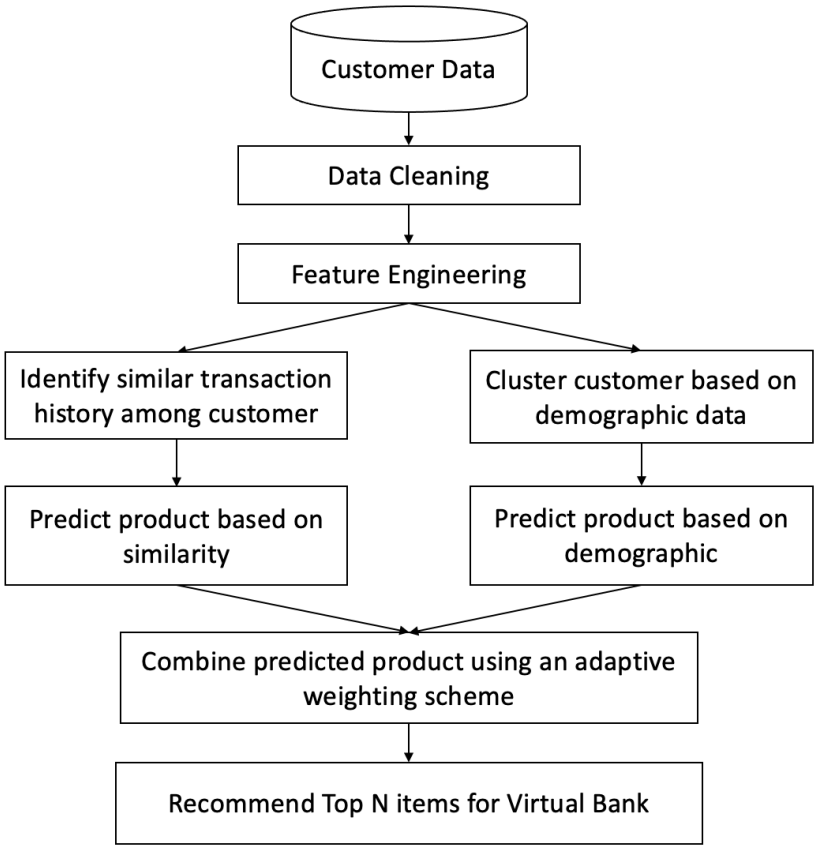
10.4.9. Tradeoff Between Recommendation Techniques

Technique	Pluses	Minuses
Collaborative Filtering (CF)	A. Can identify cross-genre niches. B. Domain knowledge nc needed. C. Adaptive: quality improves over time. D. Implicit feedback sufficient	I. New user ramp-up problem. J. New item ramp-up problem. K. "Gray Sheep" problem. L. Quality dependent on large historical dataset. M. Stability vs. plasticity problem.
Content-based Filtering (CBF)	B, C, D	I, L, M
Demographic Filtering (DF)	A, B, C	I, K, L, M N. Must gather demographic information

10.5. Customer Process Cycle



10.6. Recommendation System Architecture



10.7. General Information of Dataset

10.7.1. Number of Record in Training Set and Testing Set

```
#[Other than modeling] Check the number of records in training set and testing set

print("{} {}".format("Number of Record in Training Set: ",len(traindat)))
print("{} {}".format("Number of Record in Testing Set: ",len(testdat)))
```

```
Number of Record in Training Set:  13647309
Number of Record in Testing Set:  929615
```

10.7.2. Number of Record in Training Set by Date

```
#[Other than modeling] Summarize the counting of records in training set by date

traindat['fecha_dato'].value_counts().sort_index()
```

```
2015-01-28    625457
2015-02-28    627394
2015-03-28    629209
2015-04-28    630367
2015-05-28    631957
2015-06-28    632110
2015-07-28    829817
2015-08-28    843201
2015-09-28    865440
2015-10-28    892251
2015-11-28    906109
2015-12-28    912021
2016-01-28    916269
2016-02-28    920904
2016-03-28    925076
2016-04-28    928274
2016-05-28    931453
Name: fecha_dato, dtype: int64
```

10.7.3. Number of Record in Testing Set by Date

```
#[Other than modeling] Summarize the counting of records in testing set by date

testdat['fecha_dato'].value_counts().sort_index()
```

```
2016-06-28    929615
Name: fecha_dato, dtype: int64
```

10.8. Field Description in Dataset

Field	Data Type	Description
fecha_dato	Identifier	Record date, used to partition the dataset
ncodpers	Identifier	Customer code
ind_empleado	Demographic	Employee index (A: active, B: ex employed, F: filial, N :not employee, P: pasive)
pais_residencia	Demographic	Customer's Country residence
sexo	Demographic	Customer's sex
age	Demographic	Customer’s Age
fecha_alta	Demographic	Date in which the customer became as the first holder of a contract in the bank
ind_nuevo	Demographic	New customer Index (1: customer registered in the last 6 months)
antiguedad	Demographic	Customer seniority (in months)
indrel	Demographic	Customer type (1: First/Primary, 99: Primary customer during the month but not at the end of the month)
ult_fec_cli_1t	Demographic	Last date as primary customer (if not at the end of the month)
indrel_1mes	Demographic	Customer type at the beginning of the month (1: First/Primary customer, 2: co-owner, 3: former primary, 4: former co-owner, P: Potential)
tiprel_1mes	Demographic	Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential)
indresi	Demographic	Residence index

		(S: Yes, N: No if the residence country is the same than the bank country)
indext	Demographic	Foreigner index (S: Yes, N: No if the customer's birth country is different than the bank country)
conyuemp	Demographic	Spouse index (1 if the customer is spouse of an employee)
canal_entrada	Demographic	Channel used by the customer to join
indfall	Demographic	Deceased index. N/S
tipodom	Demographic	Addres type (1: primary address)
cod_prov	Demographic	Province code (customer's address)
nomprov	Demographic	Province name
ind_actividad_cliente	Demographic	Activity index (1, active customer; 0, inactive customer)
renta	Demographic	Gross income of the household
segmento	Demographic	Segmentation (01: VIP, 02: Individuals, 03: college graduated)
ind_ahor_fin_ult1	Product	Saving Account
ind_aval_fin_ult1	Product	Guarantees
ind_cco_fin_ult1	Product	Current Accounts
ind_cder_fin_ult1	Product	Derivada Account
ind_cno_fin_ult1	Product	Payroll Account
ind_ctju_fin_ult1	Product	Junior Account
ind_ctma_fin_ult1	Product	Más particular Account
ind_ctop_fin_ult1	Product	particular Account
ind_ctpp_fin_ult1	Product	particular Plus Account
ind_deco_fin_ult1	Product	Short-term deposits
ind_deme_fin_ult1	Product	Medium-term deposits
ind_dela_fin_ult1	Product	Long-term deposits
ind_ecue_fin_ult1	Product	e-account
ind_fond_fin_ult1	Product	Funds
ind_hip_fin_ult1	Product	Mortgage
ind_plan_fin_ult1	Product	Pensions
ind_pres_fin_ult1	Product	Loans
ind_reca_fin_ult1	Product	Taxes
ind_tjcr_fin_ult1	Product	Credit Card
ind_valo_fin_ult1	Product	Securities
ind_viv_fin_ult1	Product	Home Account
ind_nomina_ult1	Product	Payroll
ind_nom_pens_ult1	Product	Pensions
ind_recibo_ult1	Product	Direct Debit

10.9. Handling of Missing Value

10.9.1. Number of Missing Value before Data Cleaning

```
#[Other than modeling] Identify the columns with missing data

traindat.isnull().sum()
```

```
fecha_dato                   0
ncodpers                     0
ind_empleado             27734
pais_residencia          27734
sexo                     27804
age                          0
fecha_alta               27734
ind_nuevo                27734
antiguedad                   0
indrel                   27734
indrel_1mes             149781
tiprel_1mes             149781
indresi                  27734
indext                   27734
canal_entrada           186126
indfall                  27734
tipodom                  27735
cod_prov                 93591
ind_actividad_cliente    27734
renta                  2794375
segmento                189368
ind_ahor_fin_ult1            0
ind_aval_fin_ult1            0
ind_cco_fin_ult1             0
ind_cder_fin_ult1            0
ind_cno_fin_ult1             0
ind_ctju_fin_ult1            0
ind_ctma_fin_ult1            0
ind_ctop_fin_ult1            0
ind_ctpp_fin_ult1            0
ind_deco_fin_ult1            0
ind_deme_fin_ult1            0
ind_dela_fin_ult1            0
ind_ecue_fin_ult1            0
ind_fond_fin_ult1            0
ind_hip_fin_ult1             0
ind_plan_fin_ult1            0
ind_pres_fin_ult1            0
ind_reca_fin_ult1            0
ind_tjcr_fin_ult1            0
ind_valo_fin_ult1            0
ind_viv_fin_ult1             0
ind_nomina_ult1          16063
ind_nom_pens_ult1        16063
ind_recibo_ult1              0
dtype: int64
```

10.9.2. Number of Missing Value after Data Cleaning

```
#[Other than modeling] Check to make sure all missing data has been filled

traindat.isnull().sum()
```

```
fecha_dato               0
ncodpers                 0
ind_empleado             0
pais_residencia          0
sexo                     0
age                      0
fecha_alta               0
ind_nuevo                0
antiguedad               0
indrel                   0
indrel_1mes              0
tiprel_1mes              0
indresi                  0
indext                   0
canal_entrada            0
indfall                  0
tipodom                  0
cod_prov                 0
ind_actividad_cliente    0
renta                    0
segmento                 0
ind_ahor_fin_ult1        0
ind_aval_fin_ult1        0
ind_cco_fin_ult1         0
ind_cder_fin_ult1        0
ind_cno_fin_ult1         0
ind_ctju_fin_ult1        0
ind_ctma_fin_ult1        0
ind_ctop_fin_ult1        0
ind_ctpp_fin_ult1        0
ind_deco_fin_ult1        0
ind_deme_fin_ult1        0
ind_dela_fin_ult1        0
ind_ecue_fin_ult1        0
ind_fond_fin_ult1        0
ind_hip_fin_ult1         0
ind_plan_fin_ult1        0
ind_pres_fin_ult1        0
ind_reca_fin_ult1        0
ind_tjcr_fin_ult1        0
ind_valo_fin_ult1        0
ind_viv_fin_ult1         0
ind_nomina_ult1          0
ind_nom_pens_ult1        0
ind_recibo_ult1          0
dtype: int64
```

10.10. Model Performance with Different Weight between Two Filtering Approaches

```
#[Other than modeling] Check the optimal weight between memory based and demographic based

#Specify the Column Names while initializing the Table
myTable = PrettyTable(["Model", "Mixing Probability"])

model_list = ["All memory","95% memory","90% memory","85% memory", "80% memory","75% memory", "70% memory",
              "65% memory","60% memory","55% memory","50% memory","45% memory","40% memory","35% memory",
              "30% memory","25% memory","20% memory","15% memory","10% memory","5% memory","All demographics"]

memory_list = [evaluation_100, evaluation_95, evaluation_90, evaluation_85, evaluation_80, evaluation_75,
               evaluation_70, evaluation_65, evaluation_60, evaluation_55, evaluation_50, evaluation_45,
               evaluation_40, evaluation_35, evaluation_30, evaluation_25, evaluation_20, evaluation_15,
               evaluation_10, evaluation_5,evaluation_0]

for index in range(len(model_list)):
    myTable.add_row([model_list[index],memory_list[index]])

print(myTable)

max_value = max(memory_list)
max_model = model_list[memory_list.index(max_value)]

print("Maximum Mixing Probability: ", max_value, " (",max_model,")")
```

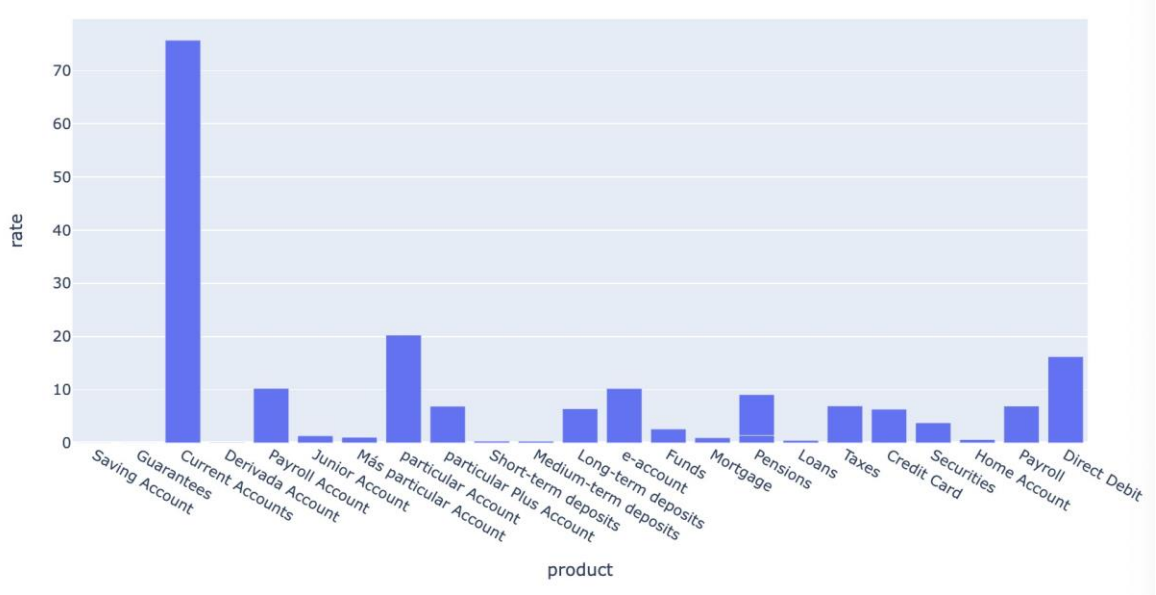
```
+------------------+-----------------------+
|      Model       |   Mixing Probability  |
+------------------+-----------------------+
|    All memory    |  0.003316739757417707 |
|    95% memory    |  0.004741021690174234 |
|    90% memory    |  0.004719292007427603 |
|    85% memory    |  0.004726535235009812 |
|    80% memory    |  0.004284039877260215 |
|    75% memory    |  0.004172099087353322 |
|    70% memory    |  0.004123371919982086 |
|    65% memory    |  0.004128639721860057 |
|    60% memory    |  0.004742338640643728 |
|    55% memory    |  0.004638299553553793 |
|    50% memory    |  0.004631714801206329 |
|    45% memory    |  0.004658712285830933 |
|    40% memory    |  0.004617886821276655 |
|    35% memory    |  0.004587596960478318 |
|    30% memory    |  0.004542820644515561 |
|    25% memory    |  0.004541503694046069 |
|    20% memory    |  0.004719950482662351 |
|    15% memory    |  0.004538211317872336 |
|    10% memory    | 0.0042122660766728505 |
|    5% memory     | 0.0044203442508527245 |
| All demographics |  0.004340668747448406 |
+------------------+-----------------------+
Maximum Mixing Probability:  0.004742338640643728  ( 60% memory )
```

10.11. Sample Report Layout

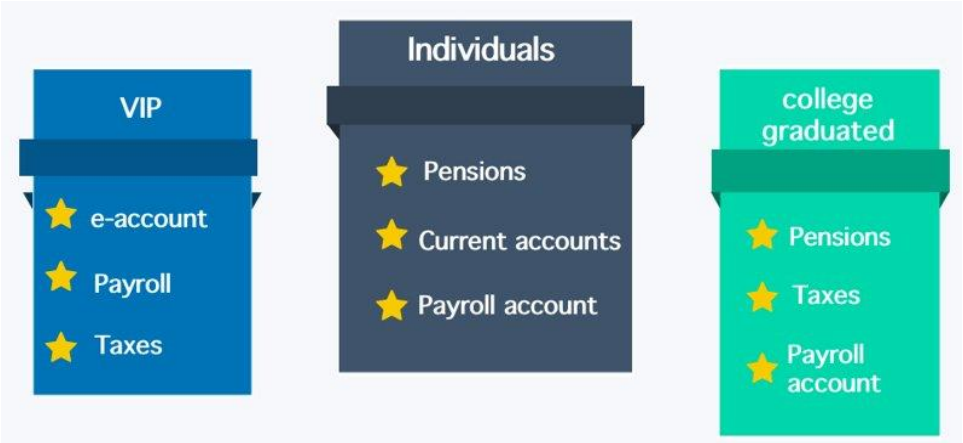
10.11.1. Customer Overview Report

10.11.1.1 Report Generated by Python

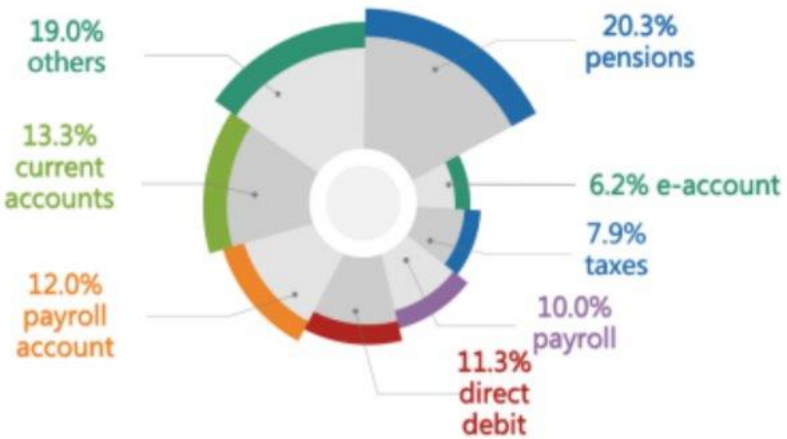
Customers' previous product usage rate



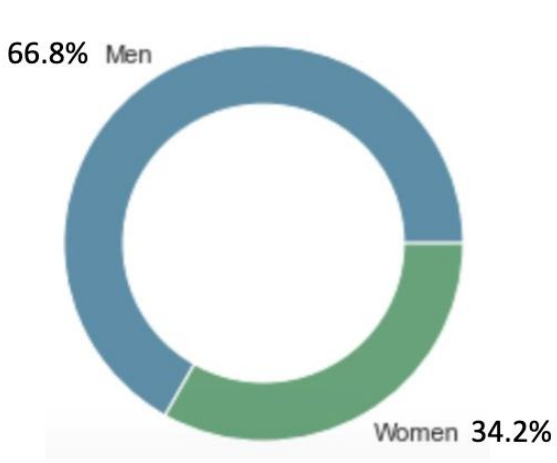
Top 3 Recommended Products for Each Customer Group

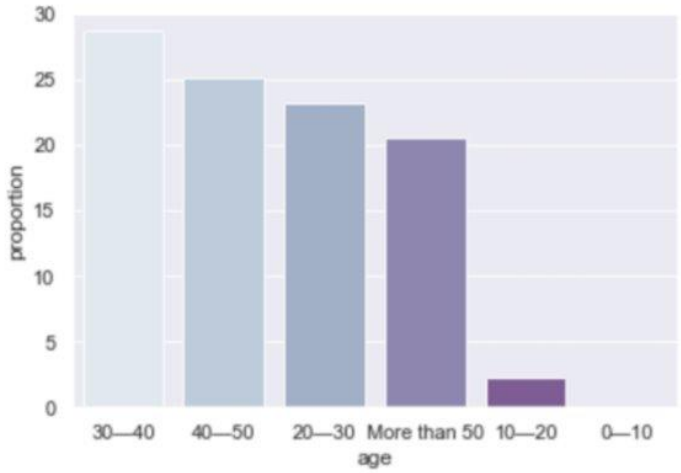


Distribution of Recommended Products



Distribution of Customer (Credit Card Product)






```
from palettable.colorbrewer.qualitative import Pastel1_7
# create data
names = ['Men', 'Women']
size = [66.8,33.2]
# Create a circle at the center of the plot
my_circle = plt.Circle( (0,0), 0.7, color='white')

plt.pie(size, labels=names, colors=["#5d8ca8","#65a479"])
p = plt.gcf()
p.gca().add_artist(my_circle)

# Show the graph
plt.show()
```

10.11.1.2 Report Generated by R Language

The report and the corresponding code are uploaded to Github (https://github.com/Maggieykw/FTEC5510-Project.git). The file names are “Customer-Overview-Report.pdf” and “Customer-Overview-Report.Rmd”. A capture of report content is shown below.

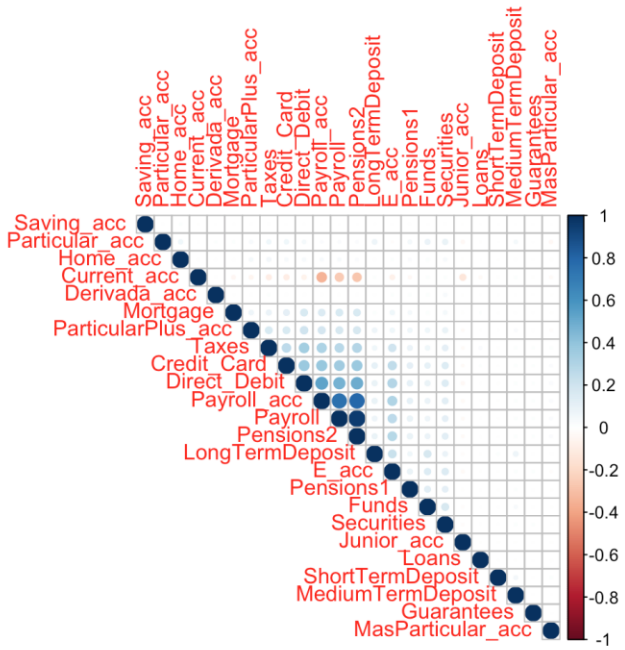
- **Report:** https://github.com/Maggieykw/FTEC5510-Project/blob/main/Customer%20Overview%20Report.pdf
- **Code:** https://github.com/Maggieykw/FTEC5510-Project/blob/main/Customer%20Overview%20Report.Rmd

1. Product Correlation

#Statistics of Product Number Purchased by Customer

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   1.000   1.474   2.000  15.000
```

#Relationship Among Product

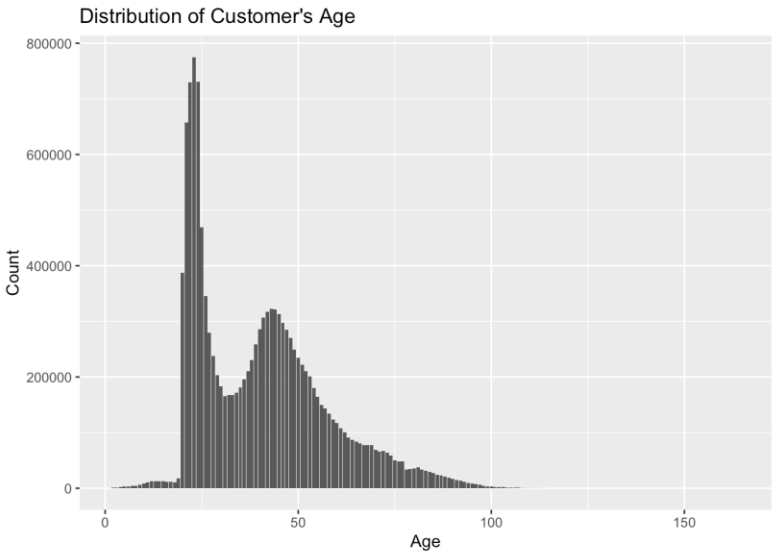


2. Age

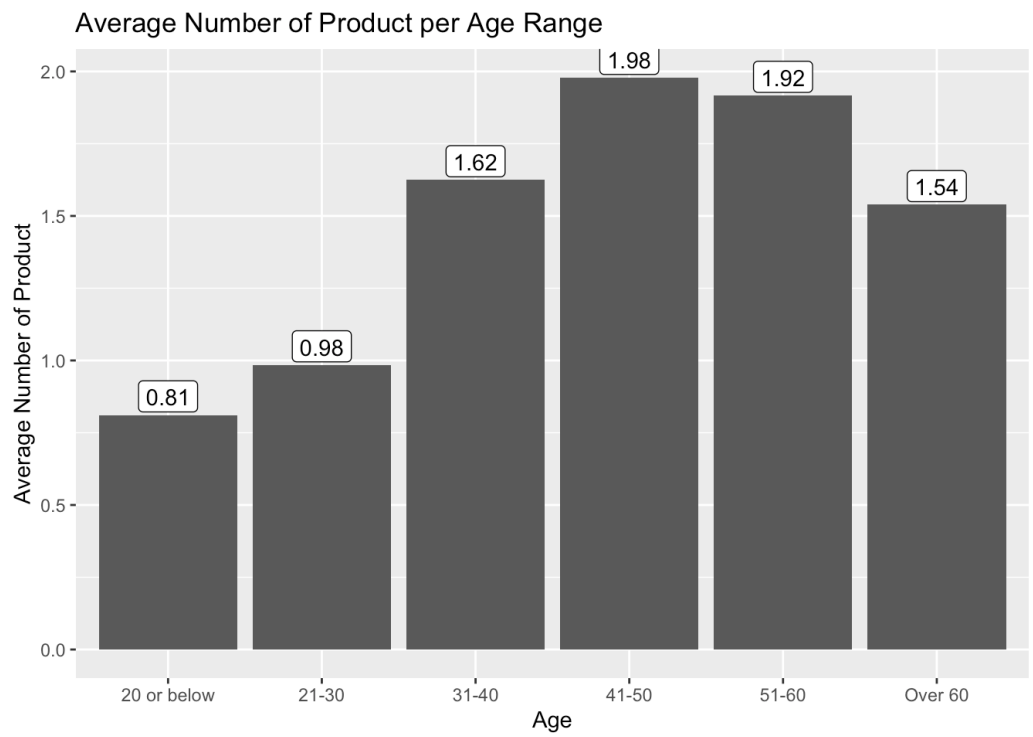
#Statistics of Customer Age

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    24.0    39.0    40.1    50.0   164.0
```

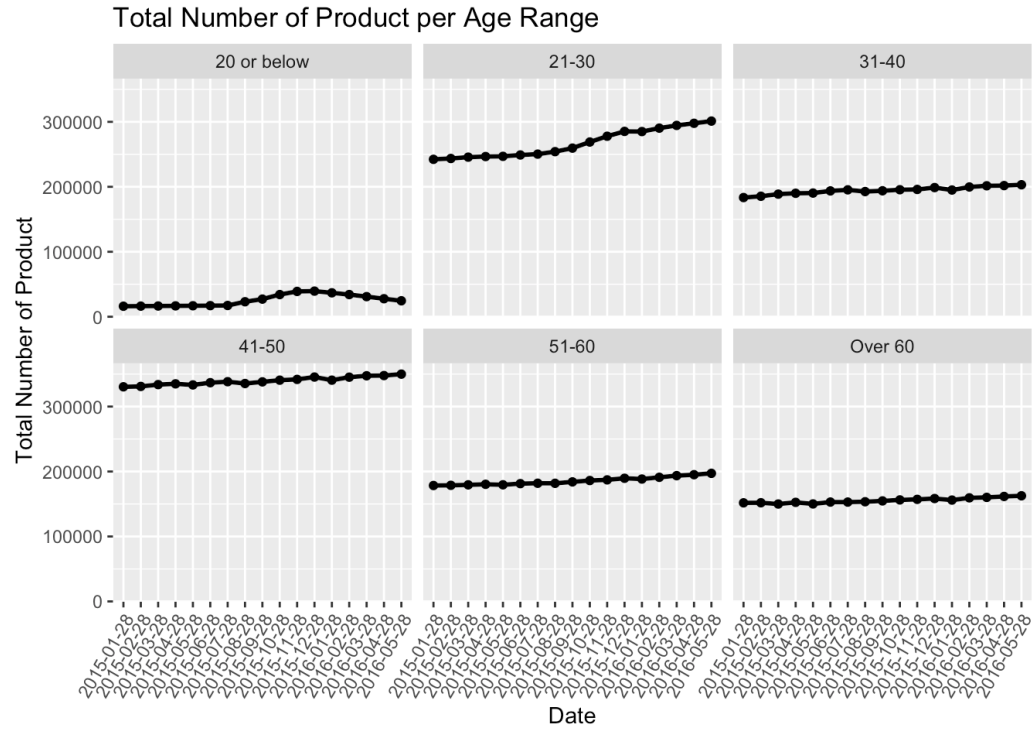
#Age Distribution



#Average Number of Product per Age



#Total Number of Product per Age



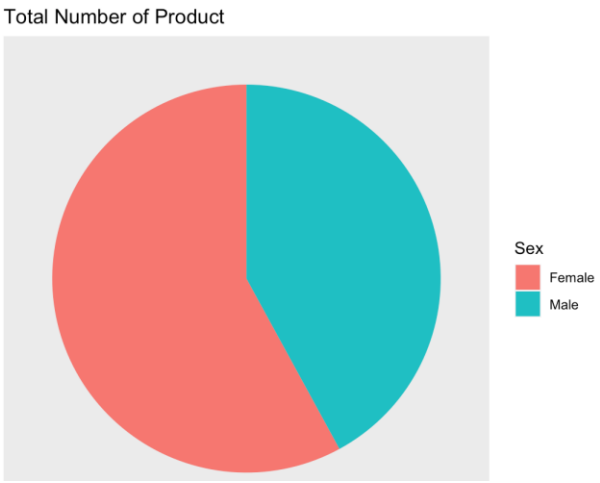
3. Gender

#Statistics of Customer Gender

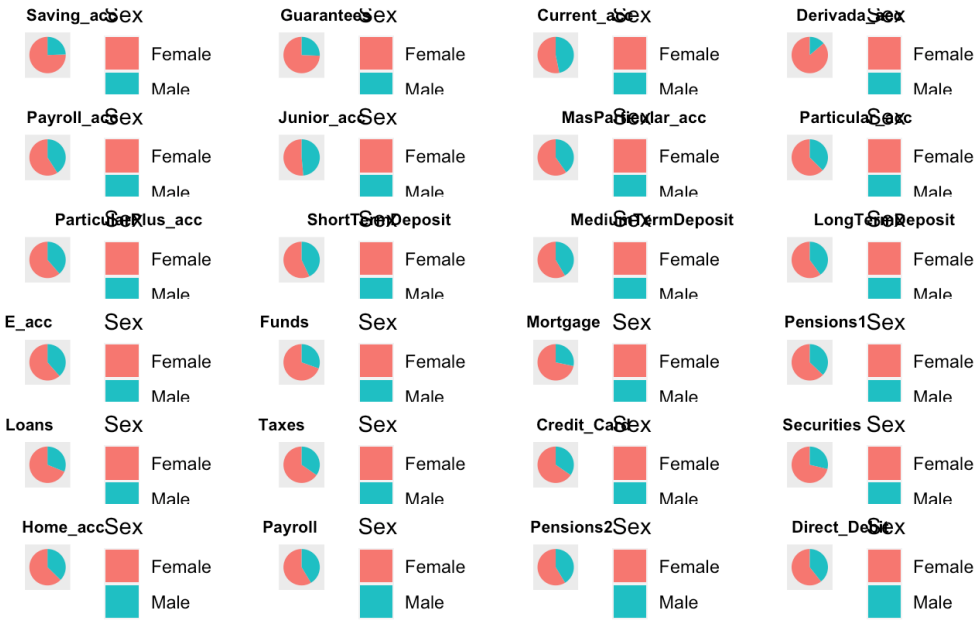
Sex <chr>	Total_Product_No <dbl>
Female	11498043
Male	8344601

2 rows

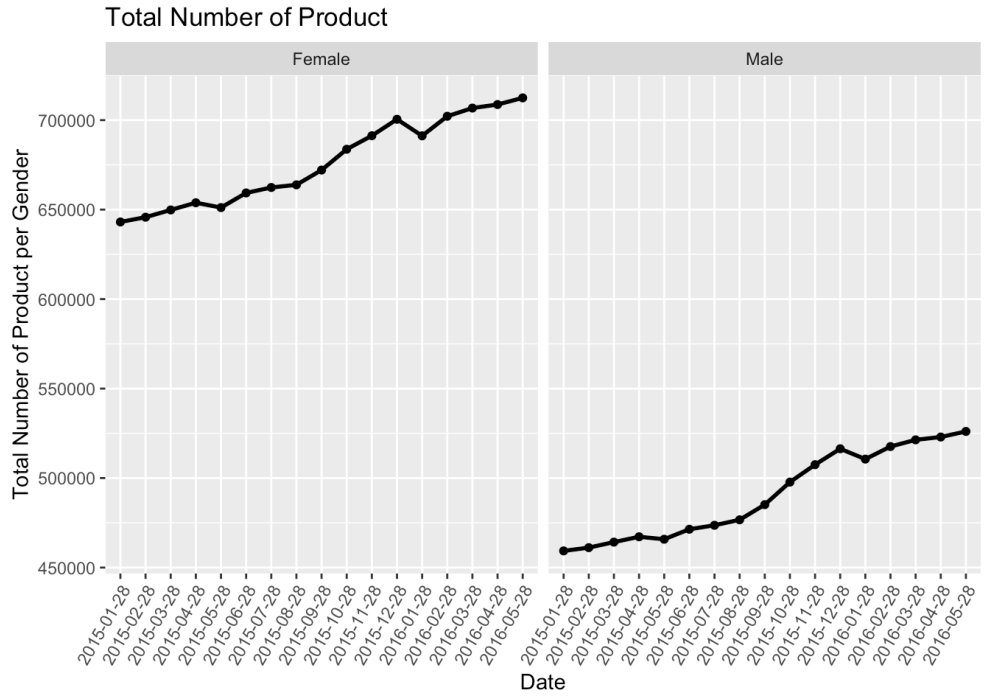
#Gender Distribution



#Age Distribution in Each Product



#Total Number of Product per Gender

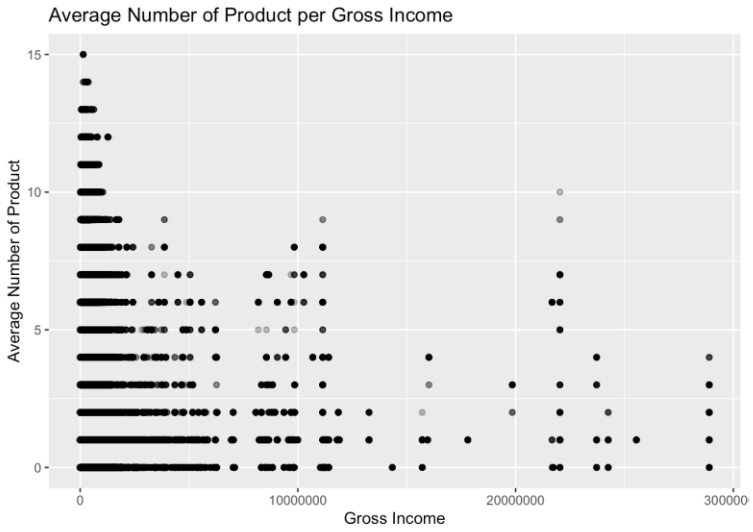


4. Gross Income

#Statistics of Customer Gross Income

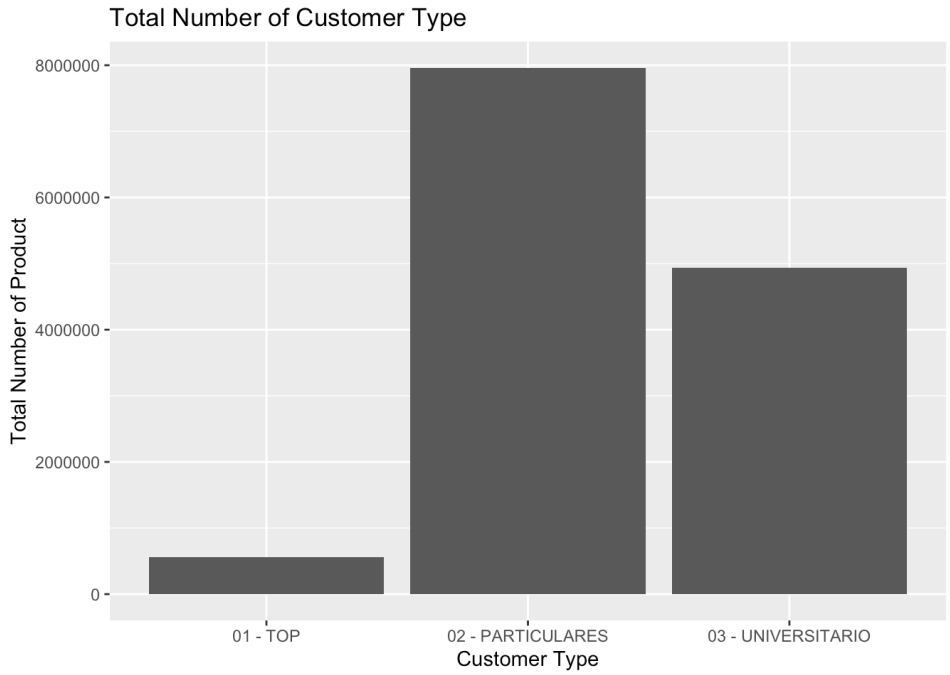
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##        0    43237    84373   107700   138148 28894396
```

#Average Number of Product versus Gross Income

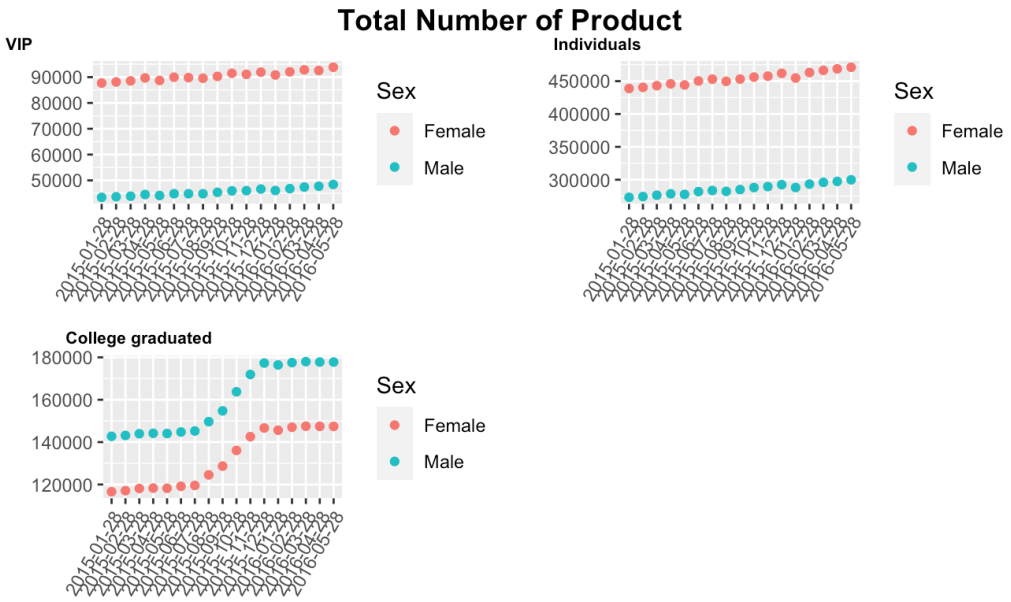


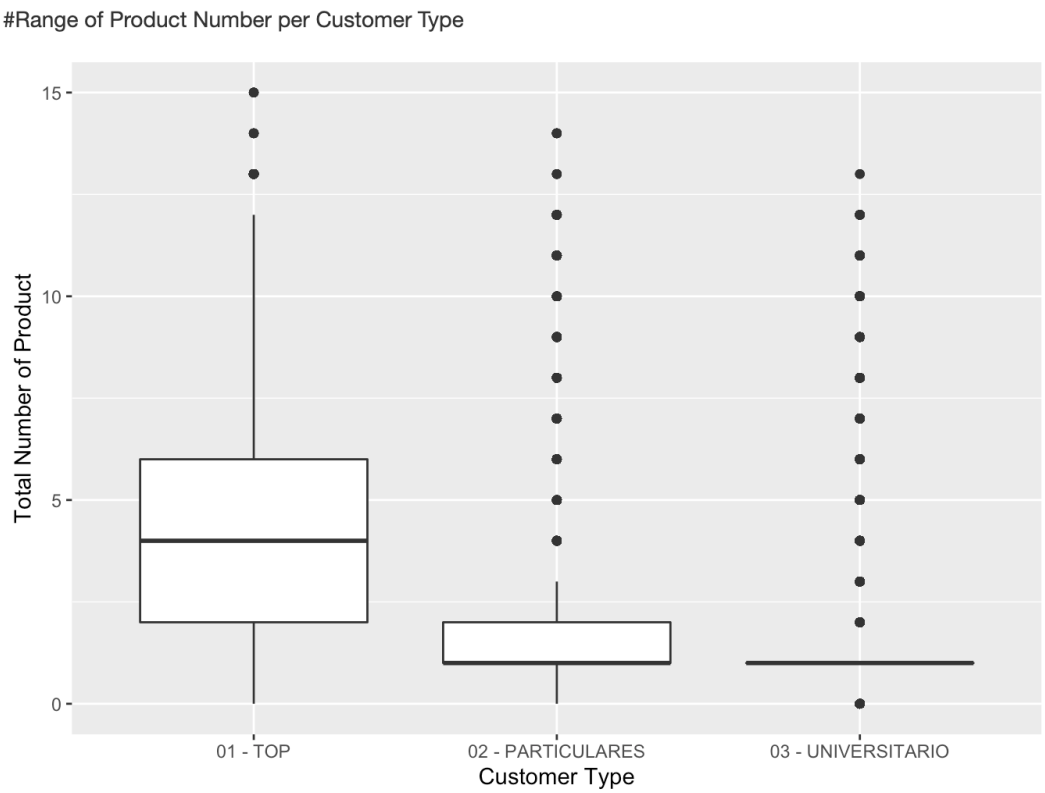
5. Customer Type

#Statistics of Customer Type



#Total Number of Product per Customer Type





10.11.2. Customer Detailed Report

Customer Report

Customer Information

Customer's ID	1384382
Gender	Male
Age	29
Residence Country	Spain
Residence Province	Leon
Foreigner or Not	N
Employee or Not	N
New Customer or Not (in 6 months)	N
Segmentation	Particular
Join Channels	KHK
Customer Seniority (in months)	9
Customer Relationship	Active
Gross Income of the Household	85725.78

Training Result

Accounts in Records	i. Payroll Account ii. Direct Debit
Marketing Direction	Credit Card

Recommendation Product

ZA Card

Card issuers	Visa
Cash Back Rate	*Shopping: up to 8%
Unique Features	Personalised Card Number; ZA Verify
https://bank.za.group/en/za-card/	

10.12. Project Timeline