## **easyMoney final project**

## **Group 3**

Hussam Almutairi

Shady A.Azeem

Bonnie Godyn

Grammer and spell check: chatgpt

## **Table of contents :**

* Introduction
* Dataset
* Analysis
* Segmentation
* Recommendation
* Monitoring
* Coordination
* Conclusion

<https://us3.ca.analytics.ibm.com/bi/?perspective=dashboard&pathRef=.my_folders%2FEasyMoney&action=view&mode=dashboard&subView=model00000185d8ef5f01_00000000>

1. access.team3.nuclio@gmail.com Pass : Team3.123456789

Code 078580

## **Introduction**

easyMoney, a finance company, is experiencing challenges related to their rapid increase of products, partner influence, and customer turnover. With a new focus on upselling to existing customers, the company has hired a data scientist to analyze their customer data and provide insights for informed decision making. This report will analyze the challenges, and provide recommendations for understanding the current customer base, aligning customer needs with business goals, creating an email campaign, and developing KPIs to enable easyMoney to become profitable again.

Task 1: Analyze customer data for demographics, product usage, retention, and channel effectiveness to identify valuable products and entry channels.

Task 2: Align customer needs with business goals by identifying attractive products and ensuring alignment with company objectives.

Task 3: Create an email campaign targeting customers with new, attractive products to increase sales.

Task 4: Develop key performance indicators (KPIs) to measure the success of initiatives and make necessary adjustments.

Task 5: Develop a timeline and allocate resources for each task to efficiently achieve the goal of upselling to the existing customer base.

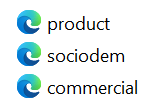
**Task 1 – Analyse the data**

## A- Datasets

There were 3 datasets provided:

* Sociodemographic\_df
* Products\_df
* Commercial\_activity\_df

Pandas profiling was used on the 3 datasets initially to understand the dataset.



After datasets were combined pandas profiling was used again to understand the data further



### Merging the datasets :



### In each dataset the pk\_cid code was used for merging the data. The pk\_cid code represents a customer number

* Upon merging each customer had multiple rows of data for each purchase, ie multiple pk\_cid codes that were duplicates. We have combined the information, so each row is an individual customer.

### Data transformation

Missing data

Data we replaced

* Segment 2.25
* Region code .04
* Entry Channel 2.23

In our machine learning model creation, we filled in missing data for the 'segment', 'region code', and 'entry channel' features using the forward fill and backfill method. This method assumes that missing values follow the same pattern as the previous or upcoming observations, depending on the fill direction. We chose this method for its simplicity, speed of application to large datasets, and usefulness when neighboring data points are highly correlated. However, it can be inaccurate if there is a rapid change in the data, which we did not detect for these features.

Data we removed

* Salary 25.36%.

We decided the salary data was too unclear to be used in our analysis. After analysis of the data it was unclear to us if the salary was per person or per household.

Date we dropped from data set

Customers who were deceased were dropped from the dataset. As the aim was to evaluate the current customer base to increase revenue and because of the extremely small number of customers this represented we decided to drop these rows.

Reducing cardinality

To simplify the dataset for analysis, we reduced the high cardinality of Age, Salary, and Entry Channel. We ensured that our categories accurately reflected the distribution of our data to minimize limitations and bias.

However, there is still a limitation as the data may change over time due to factors like shifts in marketing strategies, economic conditions, or demographic changes. To maintain accuracy and relevance, we may need to periodically update our age categories.

* Age

Age - to reduce the cardinality we created 4 different age groups

‘Minor = 0-17

Young adult = 18-30

Middle age adults = 31-65

Old aged adults = 65-110

* Salary

Salary was also categorized into income groups (low, average, above average and high). However due to the issues with missing data we did not use this feature in our project.

### Analysis the of the data:

Our analysis of the data was with respect to the business aim of increasing profitability from the current customer base. The questions we asked when analyzing the data were

* Who are our customers
* What products are the buying
* Where and how are our customers buying products

The easyMoney products can be categorized into 3 main groups and each will have a different price point. Features were created for these to be used for analysis.

Finance Products:

* Credit Card
* Loans
* Mortgage

Accounts Products:

* EM Account
* EM Account PP
* EMC Account
* EM Account P
* Payroll Account
* Debit Card

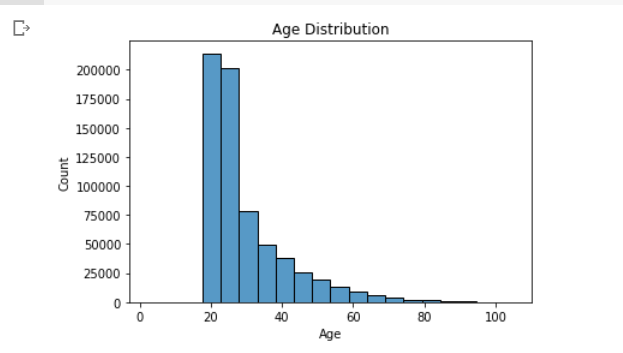
Savings Products:

* Pension Plan
* Long-term Deposit
* Securities
* Funds
* Short-term Deposit

EasyMoney, has experienced a significant decrease in revenue from 14,537,560 euros in the previous year to 3,636,970 euros in the current year (for the past 5 months). We analyze the potential reasons behind this decrease in revenue and provide recommendations to improve revenue.

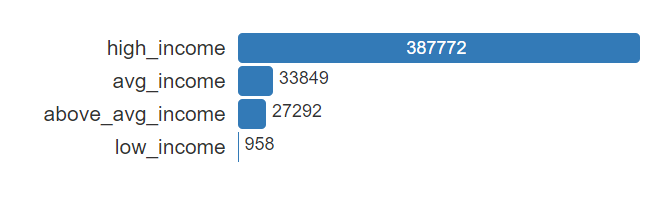
Data Analysis and Insights

Who are our customers by age

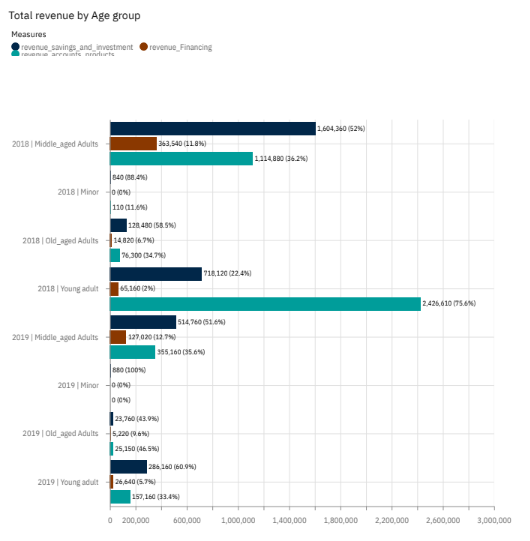


Our customers are mostly younger, around 20-30 years old. This is important information for the company as it can affect our marketing strategy.

Many of easyMoney customers are high income, however since we are unable to know if the income is a single person or everyone who lives at the address this is unfortunately not useful information for easyMoney.

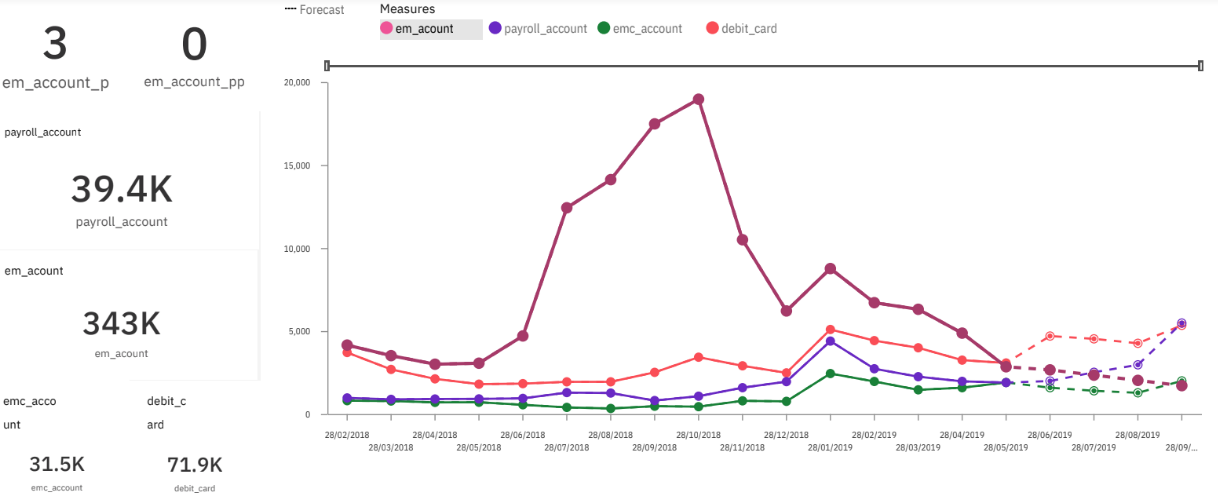


What products are they buying



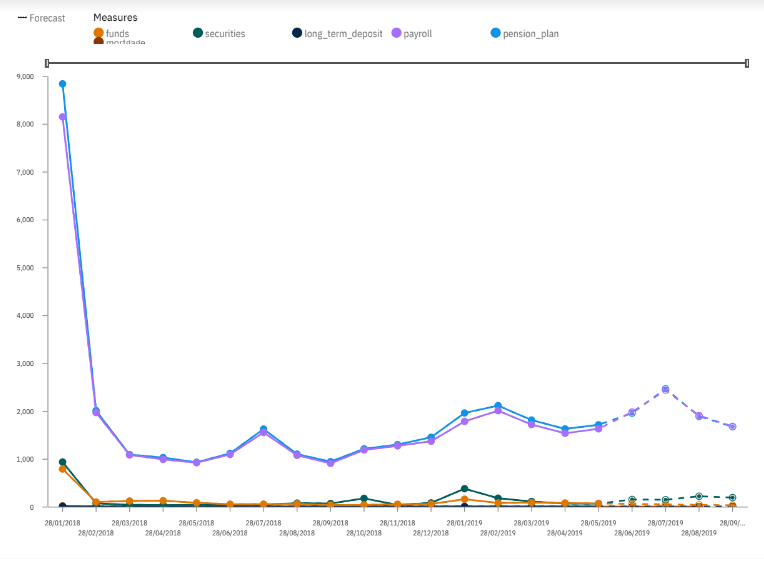
* Young adults are most attracted to account products
* Middle ages are attracted to saving an investment products
* Financing products are not strong in any age category.

Account products



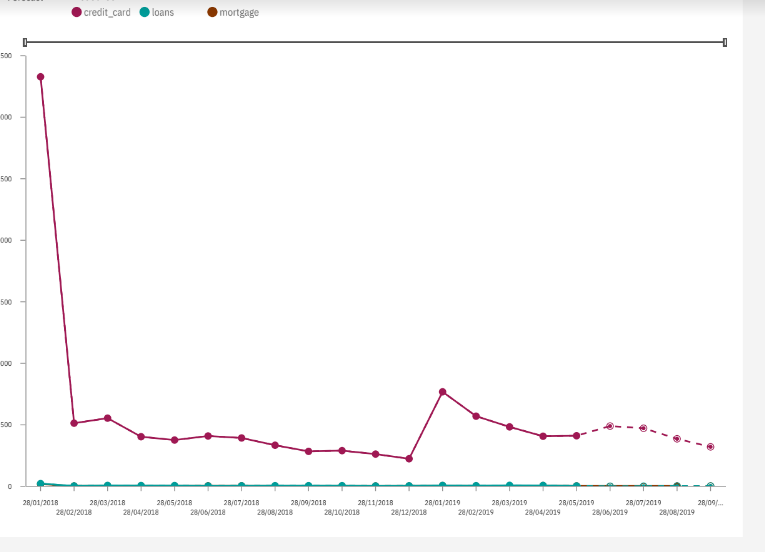
There is a small increase for the emc\_account and the debit card. In particular in January. The em-.account is decreasing.

Saving and Investment products



The initial decrease may be related to how the data was prepared - as we received data from before 2018. The payroll and pension seem to be linked accounts. A pension plan is the most popular product here and the demographics of the people buying this should be further investigated.

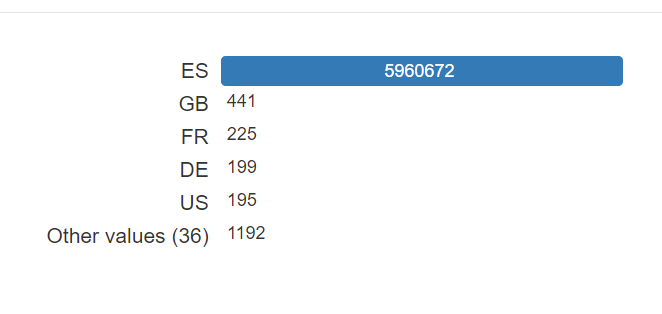
Finance



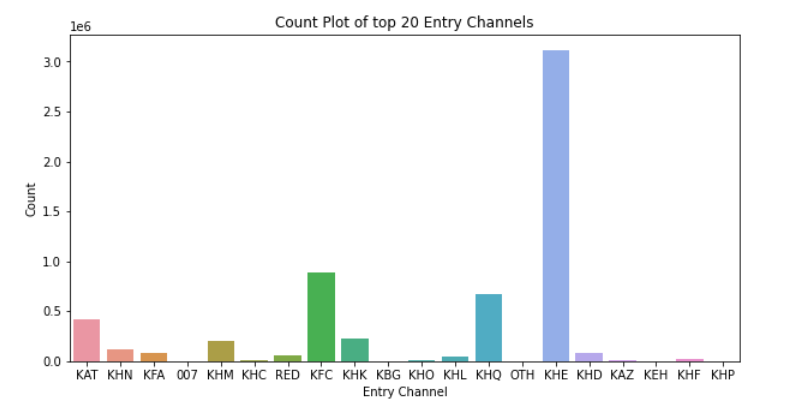
The credit card account is most popular but the loans and mortgage do not seem of interest to the easyMoney customer.

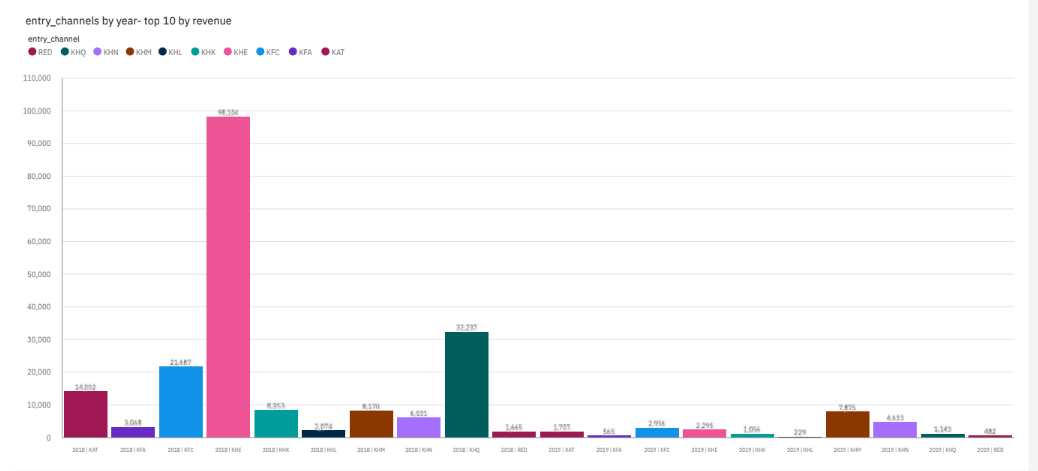
Where and how are our customers buying products

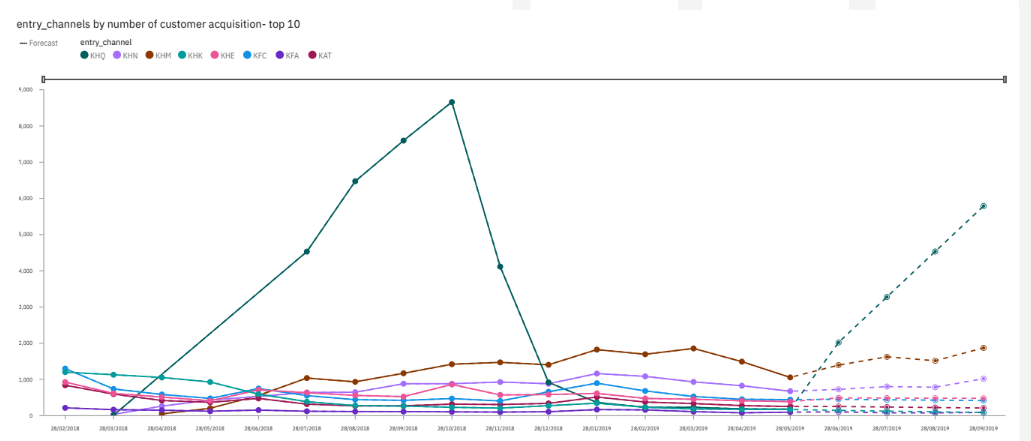
The majority of our customers are based in Spain.



What entry channels are being used and how effective are they







KHE, KFC and KHQ are contributing significantly to the revenue of easyMoney. However we can see that for many of the entry channels no or very low activity is occurring.

KHQ has a spike July - October 2019 and then a shape decrease. easyMoney could evaluate why the increase occurred and try to replicate this in other channels.

Our analysis identified the following potential reasons behind the decrease in revenue:

1. Decrease in customer acquisition: The company generated customer acquisition through only 10 out of its multiple channels, indicating that the other channels are not effectively generating customers and contributing to the decrease in revenue. easyMoney needs to evaluate the cost of these revenue channels
2. Age group of customers: Most of the revenue is generated by middle-aged adults and young adults, while the other age groups generate almost no revenue. This could indicate that the company is not effectively targeting or attracting the right age groups.
3. Low product sales: Some of the company's products, such as long-term deposits and mortgages, had very low sales between 2018 and 2019, indicating a lack of demand for these products and potentially leading to a decrease in revenue.
4. Low revenue per customer: The company is only generating 10 euros per account, 40 euros for savings and investment products, and 60 euros for financing products. If the company is not effectively monetizing its customers, this could lead to a decrease in revenue.

In terms of product sales, the accounts products sell the most, followed by debit cards and short-term deposits. Long-term deposits, mortgages, securities, and investment funds sell the least.

Recommendations

Based on our analysis, we provide the following recommendations to improve revenue:

* Evaluate marketing strategy based on the age distribution of who is buying products and who is not buying products. Eg young adults are contributing significantly to the revenue and are known to engage more with social media. Conversely older people may not have email which means that perhaps the current strategy of using an email campaign may not be effective.
* Focus on improving customer acquisition for the product segments that generate the most revenue. Currently, only 10 out of the 69 customer acquisition channels are generating revenue. Identifying the most effective channels and optimizing them could increase revenue.
* Encourage usage of payroll account services by highlighting the bonus of using the account for payroll services, instead of just using the account for payroll.
* Implement targeted marketing strategies for the young adult segment to increase revenue.
* Evaluate and optimize the most effective customer acquisition channels to increase revenue.
* Discontinue the product segments that generate little or no revenue to focus on the product segments that generate the most revenue, such as EM\_account + and EM\_account ++, as well as the securities, mortgage, and pension plan products.
* Emphasize on the product segments that generate the highest revenue: Accounts products, savings and investment products, and financing products in that order.
* Most of easyMoney customers are based in Spain, hence as they have a constrained budget it may be most practical to have all marketing material in Spanish. It may also be important that the company undertakes a competitive analysis of the Spanish market to understand what can differentiate them to the competition.

Conclusion

EasyMoney has experienced a significant decrease in revenue, potentially due to various factors such as low customer acquisition, ineffective targeting of age groups, low product sales, and low revenue per customer. Our recommendations suggest focusing on improving customer acquisition for high-revenue product segments, implementing targeted marketing strategies for the young adult segment, discontinuing low-revenue product segments, and emphasizing high-revenue product segments. Implementing these recommendations could help EasyMoney improve its revenue and profitability.

## **Task 2 Segmentation**

#### **2.1 Aim of the task :**

The management in easyMoney would like to better understand their customer base to help guide their business goals. The request is to segment customers into different groups to help guide their marketing plan. By understanding the demographics of each group, it is possible to choose the best marketing tools and strategies for each group. Different tools such as social media or email campaigns might be more appropriate for different age groups and different demographics might be attracted to different products.

Machine learning model chosen

The K-means algorithm is useful when working with unlabelled data, which means data without any pre-assigned categories or groups. This is often the case when dealing with customer segmentation data.

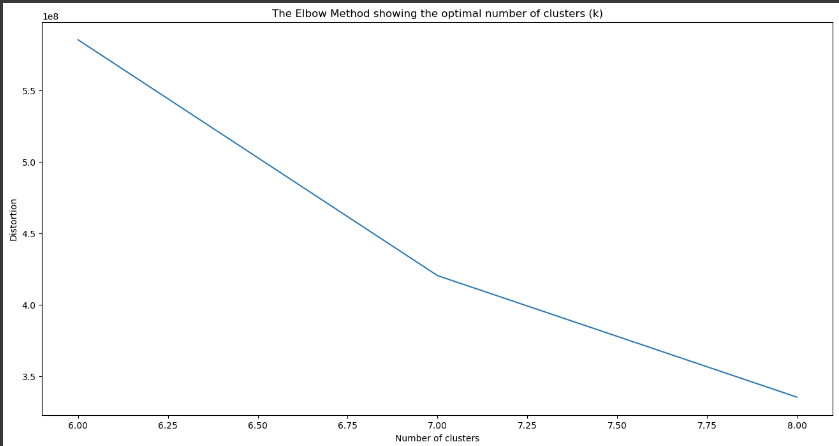
K-means works by iteratively assigning each data point to one of K clusters, where K represents the number of clusters. The algorithm discovers groups in the data based on the provided features, and is therefore quite suitable for customer segmentation.

After running the algorithm, the output consists of the centroids for each of the K clusters, as well as labels for each input data point indicating which cluster it belongs to. This results in a set of clusters that show how customers are grouped based on similarity.

Metrics for determining the optimal number of segments for easyMoney

The elbow methodfinds the value of the optimal number of clusters using the total within-cluster sum of square values. This represents how spread-apart the generated clusters are from one another. In this case, the K-means algorithm is evaluated for several values of k, and the within-cluster sum of square values is calculated for each value of k.

Accordingly, the elbow method shows that the optimal no. of Clusters is 7 .

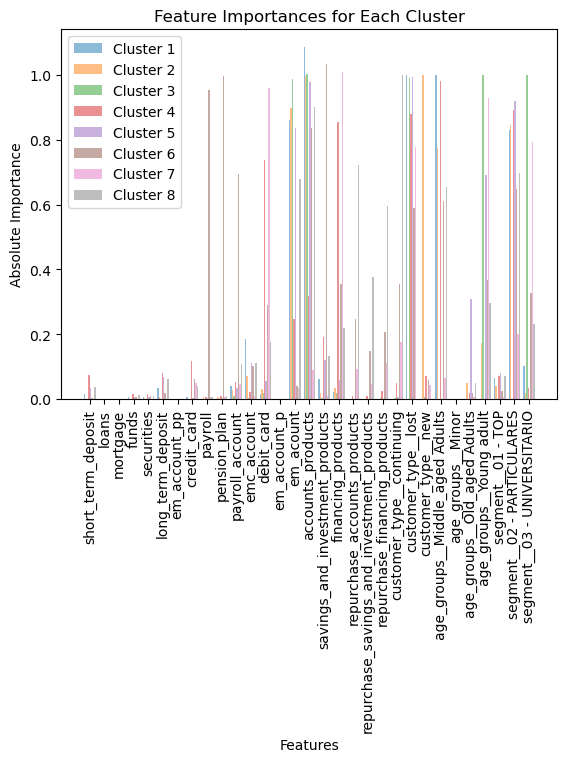


Average silhouette methodis a measure of how well each data point fits its corresponding cluster. This method evaluates the quality of clustering. As a general rule, a high average silhouette width denotes better clustering output

As it shown on the graph the optimal no. of Cluster is 8 .

## 

* Feature importance for each cluster (8 Cluster)

****

#### 2.5 Customer segmentation clusters

Cluster 1**:** Customers in this cluster are generally older (middle-aged to old-aged adults) and belong to the "02 - PARTICULARES" segment. They tend to have a high proportion of "em\_acount" products and have a high likelihood of being lost customers (customer\_type\_\_lost mean is close to 1). They don't show much interest in buying financial products, with low means for most product types.

* Generally older (middle-aged to old-aged adults)
* High proportion of "em\_acount" products
* Low interest in buying financial products

Cluster 2: Customers in this cluster are mostly new customers ("customer\_type\_\_new" mean is close to 1) and have a higher proportion of "em\_acount" products compared to other products. The mean values for most of the product types are relatively low, except for savings\_and\_investment\_products, which have a slightly higher mean value. They have varying age ranges.

* Mostly new customers
* Higher proportion of "em\_acount" products compared to other products
* Relatively low means for most product types, except for savings\_and\_investment\_products

Cluster 3: Customers in this cluster are mostly young adults who belong to the "03 - UNIVERSITARIO" segment. They have a high proportion of debit cards, indicating that they use them for everyday transactions. They also tend to have a high proportion of savings\_and\_investment\_products, which might indicate that they are saving for the future. Most of the customers in this cluster are expected to be continuing customers, with a low "customer\_type\_\_lost" mean.

* Mostly young adults who belong to the "03 - UNIVERSITARIO" segment
* High proportion of debit cards
* High proportion of savings\_and\_investment\_products

Cluster 4: Customers in this cluster have a high proportion of short\_term\_deposit, long\_term\_deposit, and funds products, which suggests that they are interested in investing and saving money. They also tend to have a relatively high proportion of securities products, which may indicate that they are interested in the stock market. Most of the customers in this cluster belong to the "02 - PARTICULARES" segment and are continuing customers, with varying age ranges.

* High proportion of short\_term\_deposit, long\_term\_deposit, and funds products
* Relatively high proportion of securities products
* Mostly continuing customers who belong to the "02 - PARTICULARES" segment

Cluster 5: Customers in this cluster have a high proportion of loans, mortgage, and financing\_products, which suggests that they are interested in borrowing money. They also have a relatively high proportion of payroll and pension\_plan products, which suggests that they may be employed and planning for their retirement. Most of the customers in this cluster are expected to be continuing customers, with a low "customer\_type\_\_lost" mean, and have varying age ranges.

* High proportion of loans, mortgage, and financing\_products
* Relatively high proportion of payroll and pension\_plan products
* Mostly continuing customers

Cluster 6**:** Customers in this cluster have a high proportion of "em\_account\_p" and "em\_account\_pp" products. They also tend to have a high proportion of "emc\_account" and "credit\_card" products. Most of the customers in this cluster belong to the "02 - PARTICULARES" segment and have varying age ranges.

* High proportion of "em\_account\_p" and "em\_account\_pp" products
* High proportion of "emc\_account" and "credit\_card" products
* Mostly customers who belong to the "02 - PARTICULARES" segment

Cluster 7**:** Customers in this cluster have a high proportion of "em\_account\_p" and "em\_account\_pp" products, i. They also tend to have a high proportion of the "emc\_account" product. Most of the customers in this cluster belong to the "01 - TOP" segment and are continuing customers, with varying age ranges.

* High proportion of "em\_account\_p" and "em\_account\_pp" products
* High proportion of "emc\_account" products
* Mostly continuing customers who belong to the "01 - TOP" segment

Cluster 8:Customers in this cluster have a high proportion of "emc\_account" and "debit\_card" products, which suggests that they use them for everyday transactions. They also tend to have a high proportion of payroll and pension\_plan products, indicating that they may be employed and planning for their retirement. Most of the customers in this cluster are expected to be continuing customers, with a low "customer\_type\_\_lost" mean, and have varying age ranges.Also this group have the highest average of repurchasing products over 70% repurchase at least 1 accounts\_products and almost 60% repurchase at least 1 financing\_products.

* High proportion of "emc\_account" and "debit\_card" products
* High proportion of payroll and pension\_plan products
* Mostly expected to be continuing customers with a low "customer\_type\_\_lost" mean.

Promising groups to focus to increase revenue

Cluster 8 seems to be a promising group to focus on. Customers in this cluster have a high proportion of "emc\_account" and "debit\_card" products, indicating that they use them for everyday transactions. They also tend to have a high proportion of payroll and pension\_plan products, suggesting that they may be employed and planning for their retirement. Most of the customers in this cluster are expected to be continuing customers, with a low "customer\_type\_\_lost" mean, and have varying age ranges. Additionally, this group has the highest average of repurchasing products, with over 70% repurchasing at least 1 accounts\_product and almost 60% repurchasing at least 1 financing\_product.

To summarize, cluster 8 is a valuable group for the business to focus on, along with clusters 4 and 5, which also show potential for investment and continued customer engagement.

Alternative models we considered

Hierarchical Clustering:

* Can be useful for visualizing how customers are related to each other.
* Difficult to determine the optimal number of clusters

DBSCAN:

* Can handle noise and outliers.
* Automatically determines the number of clusters.
* Requires setting two parameters that can greatly affect the results (minimum points and epsilon distance).

Gaussian Mixture Model (GMM):

* Can identify sub-populations that may exist within a larger customer segment.
* Can handle data that doesn't have a clearly defined cluster.
* Can be sensitive to initialization values, which can greatly affect the results.

## **Task 3 Recommendation**

Goal: Increase product sales and maximize return on investment.

Campaign: Email campaign targeting 10,000 customers with personalized product recommendations using three separate models for different price points.

* Accounts products (10 Euro)
* Saving and Investment products (40 Euro)
* Finance products (60 Euro)

To achieve this goal, three separate models were created, one for each product, and each model was optimized separately.

Strategy:

The primary objective of the email campaign is to target as many customers as possible who are likely to purchase the product. To achieve this goal, the focus was on maximizing the true positives and precision while reducing the number of false negatives.

The models were trained using historical data on customer purchases, including demographic data, purchase history, and behavior patterns. The models were then optimized to ensure that the customers with the highest probability of purchasing a specific product were targeted with the corresponding email.

Precision will be used as the main factor to evaluate model quality. This is because we want as many true positives as possible to give us the best return of investment possible. based on the overall assessment of the models on test train and validation set we will continue with the

**Features Creation for the model**

1. Rolling windows

To predict the future sales of our three product categories based on the previous sales of our two other product categories, we utilized rolling windows. Rolling windows enabled us to analyze trends and patterns in the data over time and to capture the behavior of the data in a specific period. By using rolling windows to move forward through time, we were able to identify patterns and trends that may not have been apparent in the data when viewed as a whole. This approach allowed us to create a predictive model that takes into account the historical sales of our two other product categories and their impact on the future sales of our three product categories. The use of rolling windows in this analysis provides a robust and accurate prediction of future sales and can inform our business strategy going forward.

1. Purchase sequence

Our code is used to find the purchase sequence for each segment in the data. It transforms the data by melting it into a long format, filtering out only the rows where the value is 1 (i.e., where the customer has purchased the product), and adding a purchase order for each row based on the pk\_cid, product, and pk\_partition columns.

The purchase order is then combined with the product name into a single column called 'product\_order'. This column is then used to create dummy variables for each unique combination of product and purchase order. This results in a binary matrix where each row represents a unique customer and each column represents a specific product and purchase order combination.

By analyzing this matrix, we can identify patterns in customer behavior, such as which products customers tend to purchase together and in what order. This information can be used to predict which products customers are likely to purchase next and to inform marketing campaigns and product offerings.

# entry channel

For each product group we created an entry channel count feature. The idea behind this was to count the number of channels based on the product the model was trying to predict.

* Model 1 - Account products

Notebook name : Account\_ML

Aim

Create a machine learning model to determine which customer who does not currently have a target account product would be most likely to buy an account product. One product was exempt from this - the em\_account.

* We assume most customers have an em account
* We assume an em\_account is needed to be able to have any other account products.

Modeling with and without the Payroll account was trialed, as the payroll account may not be something a customer can buy. It does not make a difference to model performance.

The model’s target (accounts target ) are

* em\_account\_pp
* emc\_account
* em\_account\_p
* Debit\_card

Feature creation

As described above, that rolling window, entry channel and a purchase sequence appropriate for the account features were created. Also, a feature that separated the customers into those who bought 2018 and earlier and after 2019 was created. The purpose of this was to be able to easily separate the data for test/train and validation.

Data split

Data was split into a 70 % training set and a 20% test set from 2018 and 2019. That means the validation data set left on 2019 around 10% to test the ability of the model’s prediction

Model

3 models were tested:

* gradient Boosting Classifier,
* Xgboost classifier°
* random Forest Classifier

The best model was Xgboost classifier° ,because comparing it with other models showed a bit higher amount of precision and accuracy on Test set as following :

XGBoosting: Accuracy: 0.88

Precision: 0,92

Gradianboosting:accuracy: 0.87

precision: 0.92

Random Forest: Train accuracy: 0.86

Train precision: 0.91

°The model was actually tuned and used hyperparameter , which are as following:

* The maximum depth of each tree is set to 6, which limits the depth of the decision trees in the model and helps prevent overfitting.
* The learning rate or eta is set to a small value of 0.0001 to ensure that the model converges slowly and carefully.
* The subsample ratio is set to 0.8, which means that each tree will use 80% of the training data, selected randomly.
* The colsample\_bytree parameter is set to 0.5, which means that 50% of the features will be selected randomly for each tree.
* The random seed used to initialize the algorithm is set to 42 for reproducibility.
* The minimum weight required to create a new node in a tree is set to 4, which controls the minimum number of instances required in each child node of a tree.
* The gamma parameter is set to 3, which controls the regularization of the model.
* The alpha and lambda parameters are set to 0.01, which are regularization parameters for L1 and L2 regularization, respectively.
* The scale\_pos\_weight parameter is set to 0.08, which adjusts the balance of positive and negative classes in the data.

Analysis

According to the dataset the percentage of customers who bought the product presents as followings

* 76,62% of customers bought a target product in 2018
* 82,51% of customers bought a target product in 2019

Based on the information provided, it seems that the ML XGB boost Classifier is performing well in terms of precision and accuracy on both the training and test data sets. The accuracy score of 0.94 indicates that the model is correctly identifying true positives and avoiding false positives, while the accuracy score of 0.85 suggests that the model is able to classify a significant proportion of instances correctly in the test set. However it must be mentioned that the model is expected to perform well on the majority class.

The use of K-Fold cross validation on the training data suggests that the model is performing well, with a mean cross-validation score of 0.92, which is a good indication that the model is stable and robust. Furthermore, achieving a mean cross-validation score of 0.94 on the Validation data set is a strong indication of the model's generalization performance.

In summary, XGB boost appears to be a good model with strong precision and accuracy scores, but further improvements may be required to enhance its generalization performance and avoid overfitting. The use of K-Fold cross validation on the training data is a good way to assess the model's stability and robustness, while achieving a high mean cross-validation score on the test data suggests good generalization performance.

Further improvement

To improve the model's performance, some potential steps include conducting additional hyperparameter optimization by experimenting with various learning rates and max depth. Additionally, increasing the size of the validation dataset might yield more representative scores, as the current dataset contains less than 10% of the total customers.

Finally, exploring feature engineering techniques could make the model predict the patterns and relationships in the data better.

* Model 2 saving and investment

Notebook – saving\_investment html

Aim

Create a machine learning model to determine which customer who does not currently have a saving and investment product would be most likely to buy a saving and investment product.

Features removed

Features that have been dropped from the model include

Payroll and any features related to this. This is because it appears the payroll account is given to anyone with a payroll service and is not an indicator that a customer will buy another product.

Any features created from the target during the ranking process – these will cause data leakage. Also removed was 2 features created as part of the process of separating customers via the year they bought a product with easyMoney.

Feature creation

As described above rolling window, entry channel and a purchase sequence appropriate for the savings and investment features was created. Additionally a feature that separated the customers into those who bought 2018 and earlier and after 2018 was created. The purpose of this was to be able to easily separate the data for test/train and validation.

Data skew and imbalance in classes

The dataset is characterized by skewness in both its features and target variables. This indicates that the distribution of values within these variables is not symmetric, with some values being more prevalent than others. Additionally, the target variable exhibits a significant class imbalance, with only around 9.5% of customers purchasing the target product in the 2018 train and test data, and 12 % of customers doing so in the validation dataset.

Models

To address this decision tree based models are used. XGBoost and gradient boosting classification models are used. On analysis the XGBoost model was overfitting. The results have been left in the notebook. Perhaps better hyperparameter tuning and feature engineering could have been used to improve the XGBoost model results.

Benefits:

* Both algorithms can effectively learn from minority class samples in imbalanced datasets.
* These algorithms can handle skewed data by partitioning the data into smaller subsets and incorporating regularization techniques to avoid overfitting.

Limitations

* These algorithms are sensitive to hyperparameter tuning, which can be challenging for users without significant machine learning experience.
* Improper hyperparameter settings can lead to overfitting and reduced model effectiveness

To address this hyperparameter optimization using a random grid search was used on the gradient boosting model. The parameters used are:

* learning rate of 0.05 to slow down the learning process and prevent overfitting.
* max\_depth=7 of the trees to allow for more complex model.
* Reduced the minimum number of samples required to split a node to allow for more splits and a more finely-tuned model. min\_samples\_split 2000.
* Increased the number of trees in the ensemble to allow for more opportunities to capture complex patterns in the data. n\_estimators= 2000.
* Reduced the subsample ratio to reduce overfitting and improve generalization performance, such as subsample=0.8
* max\_features='sqrt' was used as it prevents the model from favouring the majority class

Data split

Data was split into a 75% training set and a 25% test set, with the test set comprising customers who made their initial purchase before 2019. A validation set was also created using customers who made their first purchase after 2019 to evaluate the model's ability to generalize to unseen data, which is important for predicting future purchase behavior.

Analysis

The ML model demonstrated strong performance overall, with a particular emphasis on precision. This is reflected in the test precision score of 0.7998, which suggests that the model is effective at identifying true positives without generating too many false positives. The test accuracy score of 0.9540 also indicates that the model is able to correctly classify a high proportion of instances in the test set. However, the validation precision score of 0.7194 is noticeably lower than the training and test precision scores, which may suggest that the model is overfitting to the training and test sets and not generalizing well to new data. This is further supported by the low validation recall score of 0.0818, which indicates that the model is not effectively identifying all positive instances in the validation set. Overall, while the model performed well on the training and test data, further work may be required to improve its generalization performance. In addition to these metrics kfold validation was performed. The scores ranged from 0.947 to 0.949. The mean cross-validation score of 0.948 indicates that the model is likely to generalize well to new, unseen data.

Future improvements

* Attempt more hyper parameter optimization as some benefits were seen from changing learning rates and n estimators.
* The validation dataset was limited in data. It is just under 10% of the total customers so maybe increasing the data here might give more representation scores
* Feature engineering

**Model 3 - Finance products**

Aim

Create a machine learning model to determine which customer who does not currently have finance product would be most likely to buy a saving and investment product.

For the finance segment model, we began by creating rolling statistics and analyzing purchase sequences, as we explained in the previous chapter. We also engineered a new feature called entry\_channel\_count, where we counted the number of channels based on the finance products. We then explored various models and tuned their parameters to select our base model, as shown in the table below. We used the data from 2018 for training and 2019 data for testing, and for validation, we emphasized precision and recall in the confusion matrix to measure performance. However, unlike the other product segments, the maximum result achieved for F1 and precision was only 0.30, which is inadequate. The primary reason for this was that the data was unbalanced, with only 1.2% of the data having a target value of 1 (customers with finance products). We applied oversampling using SMOTE, but the results only improved from 17% to 30%, which is still not good. Additionally, due to time limitations, we were unable to explore other potential solutions to improve the model's performance. For example, we could have experimented with different algorithms or ensembles of algorithms that could better handle skewed and imbalanced data. We could have also revisited feature engineering and selection to identify the most significant features for predicting finance product purchases.

| Models | Confusion Matrix: | Accuracy:  Precision:  Recall:  F1-Score: | ROC | Results with grid search | Confusion Matrix with parameter from grid search |
| --- | --- | --- | --- | --- | --- |
| LOG | [[328252 1611]  [ 5101 973]] | Accuracy: 0.9800200632856756  Precision: 0.3765479876160991  Recall: 0.16019097793875536  F1-Score: 0.22476322476322477 | 0.93 | Best parameters: {'C': 0.001, 'class\_weight': 'balanced'}  Best score: 26.06%  Accuracy: 83.16% | C[[274067 55796]  [ 773 5301]]  Accuracy: 0.8316083075100391  Precision: 0.08676367088400412  Recall: 0.8727362528811327  F1-Score: 0.15783597088029058 |
| Decision Tree Classifier | NA | Accuracy: 0.92  Precision: 0.09  Recall: 0.34  F1-Score: 0.14  AUC: 0.64 | 0.73 | Best parameters: {'max\_depth': 7, 'max\_features': 'sqrt', 'min\_samples\_leaf': 1, 'min\_samples\_split': 5}  Best AUC score: 0.9251882115504525 | Accuracy: 98.00%  Precision: 0.37  Recall: 0.15  F1-Score: 0.22  AUC: 0.92  Confusion Matrix:  [[328267 1596]  [ 5133 941]] |
| Random Forest | Confusion Matrix:  [[327272 2591]  [ 4608 1466]] | Accuracy: 0.98  Precision: 0.36  Recall: 0.24  F1-Score: 0.29  AUC: 0.62 | 0.85 | Best parameters: {'bootstrap': True, 'max\_depth': None, 'max\_features': 'log2', 'min\_samples\_leaf': 4, 'min\_samples\_split': 9, 'n\_estimators': 530}  Best AUC score: 0.6335351966873706 | Best parameters: {'bootstrap': True, 'max\_depth': None, 'max\_features': 'log2', 'min\_samples\_leaf': 4, 'min\_samples\_split': 9, 'n\_estimators': 530}  Best AUC score: 0.6335351966873706  Accuracy: 0.98  Precision: 0.40  Recall: 0.12  F1-Score: 0.19  AUC: 0.56  Confusion Matrix:  [[32951 108]  [ 522 73]] |
| Random Forest with tuning based on grid search | Confusion Matrix:  [[328754 1109]  [ 5224 850]] | Accuracy: 0.98  Precision: 0.43  Recall: 0.14  F1-Score: 0.21  AUC: 0.57 | 0.94 | NA | NA |
| KNN model | onfusion Matrix:  [[33059 0]  [ 595 0]] | Accuracy: 0.98  Precision: 0.00  Recall: 0.00  F1-Score: 0.00  AUC: 0.50 |  | KNeighborsClassifier(metric='manhattan', n\_neighbors=9) | NA |
| KNN model with over sampling | Confusion Matrix:  [[96521 446]  [ 1676 377]] | Accuracy: 0.98  Precision: 0.46  Recall: 0.18  F1-Score: 0.26  AUC: 0.59 | 0.59 | NA | NA |
| XGBoost model | Confusion Matrix:  [[329854 9]  [ 6062 12]] | Accuracy: 0.98  Precision: 0.57  Recall: 0.00  F1-Score: 0.00  AUC: 0.93 | 0.93 | NA | NA |
| XGBoost model\_ best results | Confusion Matrix:  [[95688 1279]  [ 1282 771]] | Accuracy: 0.97  Precision: 0.38  Recall: 0.38  F1-Score: 0.38  AUC: 0.94 | 0.94 | NA | NA |
| ## Model 6: Support Vector Machines | Confusion Matrix:  [[95688 15873]  [ 2773 3301]] | precision recall f1-score support  0 0.99 0.95 0.97 329863  1 0.17 0.54 0.26 6074  accuracy 0.94 335937  macro avg 0.58 0.75 0.62 335937  weighted avg 0.98 0.94 0.96 335937 | 0.89 | NA | NA |
| ## Model 7: Naive Bayes classification | Confusion Matrix:  [[89447 7520]  [ 667 1386]] | Accuracy: 0.92  Precision: 0.16  Recall: 0.68  F1-Score: 0.25  AUC: 0.80 | NA | NA | NA |

**Measuring the return on investment**

To evaluate the return on investment for the campaign, we propose modifying the precision score to account for the revenue generated by each product and selecting the product that provides the highest return. This approach involves assigning weights to each product based on their respective prices, with account products weighted at 1, savings and investment products weighted at 4, and finance products weighted at 6. We can then modify the precision score to reflect these weights, for example, by calculating precision as (4 \* true positive predictions for savings and investment) / (4 \* total positive predictions for savings and investment + 1 \* false positive predictions for savings and investment). Using this modified precision score, we can estimate the expected return on investment for each product at various decision thresholds. By selecting the threshold that maximizes the expected return on investment, we can determine the optimal strategy for the campaign. If the expected ROI for savings and investment is higher, the threshold for savings and investment is chosen, otherwise, the threshold for finance products is chosen. This approach assumes that the modified precision score is a good indicator of the effectiveness of the campaign for each product category, and that maximizing expected ROI is the main objective of the campaign.

An example code is

# Define product weights

account\_weight = 1

savings\_weight = 4

finance\_weight = 6

# Calculate modified precision score for savings and investment

precision\_si = (tp\_si \* savings\_weight) / ((tp\_si \* savings\_weight) + (fp\_si \* account\_weight))

# Calculate modified precision score for finance

precision\_f = (tp\_f \* finance\_weight) / ((tp\_f \* finance\_weight) + (fp\_f \* account\_weight))

# Calculate expected ROI for savings and investment

expected\_roi\_si = precision\_si \* savings\_price

# Calculate expected ROI for finance

expected\_roi\_f = precision\_f \* finance\_price

# Choose threshold that maximizes expected ROI for the campaign

if expected\_roi\_si > expected\_roi\_f:

threshold = threshold\_si

expected\_roi = expected\_roi\_si

else:

threshold = threshold\_f

expected\_roi = expected\_roi\_f

Limitations of this approach are

* There are three different models that are optimized differently and trained on slightly different data, which may introduce inconsistencies or errors when trying to compare and combine their results.
* The weights assigned to each product category may not fully capture the value or profitability of each individual product within that category, which could result in suboptimal decisions or missed opportunities for generating revenue.
* The precision score may not fully capture the complexity of the decision-making process, as it only considers the ratio of true positives to the total number of positive predictions and does not take into account other important factors such as customer behavior or market trends.
* Additional analysis and validation may be needed to ensure that the approach is robust and effective in achieving the desired results.

## **4 Monitoring**

KPIs are important for measuring the success of a marketing campaign as they help businesses track progress towards specific goals and objectives. When targeting current customers for different products, KPIs can provide valuable insights into customer behavior and preferences, helping businesses refine their marketing strategies and improve overall ROI

1. "Demographic Engagement Analysis":

Identification of groups/subgroups based on demographic information that are not engaging with targeted marketing. A decision can be made to adjust marketing tactics or to focus only on groups that are engaging, based on potential revenue.

Average Revenue per Subscriber = (Total Revenue Generated / Number of Subscribers) \* 100

1. "Product Performance Analysis":

Identification of areas or features where the company is not achieving its goals, in order to reduce costs. Customers were given suggested products based on modeling, and management wants to understand what these products are. Further analysis on market needs, competitive analysis, etc. should be performed to understand whether the product should be developed to meet customer needs or be end-of-life.

Engagement Rate by Demographic Group/Subgroup = (Number of Engagements / Total Number of Contacts) \* 100

1. "Product Revenue Growth":

To improve profitability, the company wants to grow revenue by X percent for each product. This can be measured by looking at the percentage of customers who were contacted about a product and subsequently made a purchase, as well as the percentage of customers who were contacted about a product and then purchased another product.

Product Performance Index = (Total Sales Revenue of the Product / Total Cost to Produce the Product) \* 100

The definition of the percentage would need to be made based on many factors, such as cost to develop and maintain product, profit of product and cost of marketing for a particular product.

1. Product Revenue Growth Rate

Measure the increase in product revenue after the email marketing campaign.

(Total Revenue of the Product after Marketing Efforts - Total Revenue of the Product before Marketing Efforts) / Total Revenue of the Product before Marketing Efforts \* 100

## **ML ops**

For easyMoney to continuously grow revenue they will need to monitor and retrain their machine learning models. The models we created for the email campaign for example could be applied to new customers or to existing customers to encourage them to buy more products in the future. Having a proper infrastructure to deploy, monitor and manage machine learning models in product is critical for the company's success. However they need to balance this with the cost and complexity of the solution they choose

Once the models are deployed the following could be implemented

1. Model deployment: This stage involves deploying the trained model to a production environment, where it can be used to make predictions on new data.
2. Model monitoring: This stage involves monitoring the performance of the deployed model in production, to ensure that it is working as expected and that any issues are quickly identified and addressed.
3. Model management: This stage involves managing the various versions of the model and tracking changes to the code and data over time.
4. Model optimization: This stage involves improving the performance of the model over time by tuning hyperparameters, retraining with new data, and making other adjustments as needed

ML ops tools such as GIT, Docker, Kubeflow, TensorFlow Extended, Elasticsearch and MLFlow could be utilised by easyMoney.

## **5 Coordination**

Project plan and timelines can be found here



**Conclusion**

easyMoney’s customers are younger and tend to prefer account, finance and then saving and investment products. Their best customer segment has a high proportion of "emc\_account" and "debit\_card" products, indicating that they use them for everyday transactions. They also tend to have a high proportion of payroll and pension\_plan products, suggesting that they may be employed and planning for their retirement. The products of funds and securities are not appreciated by customers and may need to be dropped. There are many acquisition channels for the company but unfortunately many do not generate revenue. An email campaign will be launched to encourage customers to buy additional products. Continued investment in the data science team will help easyMoney continue to understand their customers and be able to take advantage of this knowledge to generate future sales.