**Final Project: Rosetta Stone Subscriber Analysis**

Chapman university | 30 November 2023

MGSC 410: Applied Business Analytics

David Veksler, Willie Sine, Kyle Sharp, Kreesh Teli,

Samith Lakka, Matthew Urata-Espinosa

**Executive Summary:**

This report presents the modeling, analysis, and insights derived from examining the data frame concerning Rosetta Stone subscribers. Each objective presented distinct challenges, necessitating specialized data manipulation and model application to fulfill.

To identify customer segments, a Tableau dashboard was employed for an in-depth analysis of the data frame. Segmentation was conducted using a percentile-based method, resulting in the categorization of four distinct segments: undesired subscriber, casual subscriber, cash flow subscriber, and desired subscriber.

The identification of the most valuable subscribers revealed that those with extended subscription durations and substantial purchase amounts were of the highest value. These subscribers typically chose auto-renewal, exhibited high ID counts, were often lifetime subscribers, and demonstrated a high email open rate. This conclusion was supported by data from percentile-based segmentation and insights from random forest regression analysis.

In accentuating subscribers with upselling potential, it became evident that a K-Means clustering algorithm would enable the ability to identify these desired customer segments. After tuning the hyperparameters and building the finalized model, four clusters emerged. The two customer bases identified as being highly attractive from the model (‘Loyal Subscribers’ & ‘Life-Time Subscribers’) had comparatively high amount of email engagement, high propensity for auto renew to be on, and fairly high subscription duration and purchase amount. Should Rosetta Stone gather additional customer information, they should aim to market additional products and services to these respective customer groups.

To identify indications of subscriber churn and enable the ability to predict this churn, two classification models were implemented. The first, a logistic regression trained to predict ‘Auto Renew’, this model quantified that the biggest indications of churn subscription duration and number of emails sent. A random forest classification model was built to supplement the findings from the logistic regression. While both model’s indicated similar feature importance, the more complex random forest model enables Rosetta Stone executives to run new customers through this model and predict if they will churn at the end of their billing cycle with 84% accuracy.

When identifying additional business opportunities for Rosetta Stone, the focus was on analyzing variables such as ID Count, lead platforms, user types, and push notifications. Leveraging subscriber data for monetization and strategic use was highlighted, indicating that detailed analysis can provide valuable insights for various entities. Implementing gamification elements like quests and challenges, linked with push notifications, aimed to boost user engagement and satisfaction. The strategies also included tailoring offerings to the larger consumer base, suggesting features such as in-app transactions and premium badges to incentivize spending, thus retaining and monetizing existing customers. These approaches are designed to contribute to Rosetta Stone's overall growth and profitability, enriching the user experience and encouraging ongoing engagement and financial investment.

**Analytical Plan:**

The following breakdown summarizes key steps taken in the completion of this report.

1. **Preliminary Evaluation:** Calculated mins and maxes of each data column, summary statistics of each feature in the dataset, created requirements for cleaning.
2. **Data Cleaning and Merging:** Conducted thorough data cleaning for modeling, summarizing key app activity data and integrating it into the subscriber dataset, converting purchase amounts to USD, removing outliers, eliminating unusable columns, created exploratory data visualizations of the dataset to get a grasp of the data’s distribution.
3. **Initial Modeling:** Performed linear and logistic regressions, clustering analysis, random forest regression, and random forest classification analysis to develop models that yield critical insights aligned with project goals, including identifying valuable subscribers, key segments, and accentuates prospective customers to sell additional products and services to.
4. **Analysis with Initial Modeling:** Utilized developed models to inform and refine project strategies and identified knowledge gaps to be filled with additional modeling.
5. **Additional Modeling:** Adjusted or created additional models to fill knowledge gaps, aiming to yield conclusive findings for each project goal.
6. **Conclusive Analysis:** Modeling insights revealed clear and definitive results, which were summarized in a report and group presentation.

**Action Plan Breakdown:**

This report outlines the steps we followed to thoroughly understand the complete objective:

1. Data Cleaning and Merging – In order to successfully construct the models necessary for our analysis plan, it became evidently clear that there was an inherent need to clean and merge the datasets. To begin with this, we conducted an evaluation of the datasets and narrowed down which columns could pertain significance to the models we were planning to build. First, we identified a need to alter certain aspects of the datasets like standardizing the currency amount and creating a purchase cap amount. After this, a new feature, ‘Email Open %’, was created in order to better handle some aspects of the data. Lastly, general cleaning was performed on the datasets and unnecessary columns were dropped, along with a proper review and transformation of certain data types.
2. Modeling Analysis and Techniques – We employed a multitude of models to draw insights from in pursuit of determining the most valuable subscribers. Five different models were constructed to address the objectives tasked; a linear regression model, a random forest classification model, two random forest regression models, a K-Means clustering algorithm, a percentile-based segmentation, and a logistic regression model. Additionally, we employed a variety of differing analysis techniques to generate conclusions to provide explanations for the intended objectives outlined in the project.
3. Addressing Objectives – In order to address the objectives tasked, each group member focused upon constructing models and conducting analysis that provided evidence which supports the group's dissertation provided in this report. Each of the models constructed produced adequate and insightful results which are detailed in the report below.
4. Division of Work Across Group – The division of Work Across Group section below details each group member's individual contributions to the project. To create a strategic plan of action, there was an essential need for proper communication amongst group members so various group meetings outside of class were scheduled to further discuss and complete the necessary work. More specific details of the division of work can be found below.

**Division of Work Across Group:**

* David Veksler cleaned and merged data, created linear regression model and percentile-based segmentation analysis, as well as identified the most valuable subscribers', aided in action plan development and template design.
* Willie Sine abstracted and built both random forest regression models, the random forest classification model, K-Means clustering algorithm, as well as the identifying customers likely to be upsold section, and assisted with the subscriber churn section.
* Kyle Sharp constructed the logistic regression model, variables leading to higher likelihood of auto renewal, as well as identified subscriber churning and other possible solutions.
* Kreesh Teli developed the subscriber segmentation and worked on the business opportunities section.
* Matthew Urata-Espinosa helped to develop the business opportunities section, developed the action plan, and completed the division of work across group section.
* Samith Lakka tested an alternative option for a model and helped throughout the project.

**Data Cleaning and Merging:**

In its raw form, the original datasets provided would have posed significant challenges for any modeling intended to yield consequential results. Following the preliminary evaluation of the subscriber and app activity files, the subsequent step involved thoroughly cleaning this data into a unified and accessible data frame. This was crucial to avoid additional complications in deriving insights from future modeling efforts. The initial subscriber information dataset had row-to-column dimensions of 24:40,013. In contrast, the newly cleaned and merged data frame featured dimensions of 27:10,043, as also detailed in the table below. The updated data dictionary for the updated data frame can be found in Appendix 1.1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Pre-Cleaning Dimensions** | | **Post-Cleaning Dimensions** | |
| **Rows** | **Columns** | **Rows** | **Columns** |
| **24** | **40,013** | **27** | **10,043** |

The data frame was transformed into a usable form through a series of actions, listed here in order of importance, from the most to the least critical:

**Currency Standardization**:

* Converted all purchase amounts to USD using recent exchange rates for various currencies. Refer to Appendix 1.2 for rates.

**Purchase Amount Cap**:

* Eliminated all rows with values of 1,000 or greater for their currency-adjusted purchase amounts to prevent outliers from skewing modeling results.

**Feature Engineering**:

* Created a new feature, 'Subscription Duration', representing the number of days between subscription start and end dates.
* Created a new feature, ‘Email Open %’, as a way to track each customer’s engagement with Rosetta Stone’s marketing team.
* Created new feature, ‘Email Impression %’, representing another way to track customer engagement.

**Missing Data Handling**:

* Dropped columns with a high percentage of missing values and imputed missing values in other columns where appropriate.

**Removing Unnecessary Columns**:

* Dropped columns with only one unique value.

**Outlier Detection and Treatment**:

* Identified and treated additional outliers in numerical columns using Interquartile Range (IQR) method.

**Data Consistency Checks**:

* Ensured logical consistency, such as subscription end dates being after start dates.

**Data Type Corrections**:

* Converted columns with yes-no data options to binary dummy values for modeling and converted date columns to datetime format and ensured numeric columns were correctly typed.

**Modeling and Analysis Techniques:**

A variety of models and quantitative analysis methods were employed to derive insights that were instrumental in fulfilling the project objectives. While some models were initially developed to establish baseline metrics, these and additional models were later refined to enhance insights and provide more definitive evidence in areas where knowledge gaps were prevalent. As outlined in the descriptions of each model, every modeling attempt had its limitations and varied in utility. However, through these imperfect results, additional insights about the data were uncovered.

**Logistic Regression:**

Logistic regression was utilized to predict the factors that most influenced a subscriber's decision to auto-renew their service. Analyzing the model's odds coefficients provides insights into the significance of each feature, such as email engagement or trial usage, in affecting renewal odds. These insights are instrumental in forecasting subscriber behavior, informing targeted marketing, and enhancing customer retention strategies. Furthermore, they play a key role in guiding product development and strategic decision-making, enabling a tailored approach to meet subscriber needs and effectively drive business growth.

The study on subscriber auto-renewal decisions revealed that being a 'Demo User', 'Free Trial User', or 'Email Subscriber' has a positive effect on the likelihood of auto-renewal, indicating that engaging trial experiences and regular email communication contribute to fostering subscriber loyalty. On the other hand, a high 'Send Count' tends to decrease the chances of renewal, suggesting a risk of communication fatigue. Additionally, active engagement, as evidenced by a high 'Open Count', is associated with increased auto-renewal rates, whereas 'Click Count' appears to have a negligible effect. An interesting observation is that a longer 'Subscription Duration' marginally lowers the likelihood of renewal, possibly reflecting a subscriber's reevaluation of the service over time. Regional differences, such as those represented by 'Country\_Europe', 'Country\_Other', and 'Country\_US/Canada', highlight the impact of local market and cultural influences. Understanding these aspects is vital not just for comprehending auto-renewal trends but also for pinpointing subscribers who are most likely to be receptive to additional products or services. Subscribers who actively engage, especially those responsive to trials and consistent updates, stand out as key targets for upselling and cross-selling initiatives, given their propensity for continued and expanded service usage.

**Linear Regression:**

Given Rosetta Stone’s position as a business intent on reaching profitability and maximizing returns from operations, comprehending the trends that affect the total purchase amount per customer was crucial to achieving the project’s objectives. A linear regression model was developed to gain a deeper understanding of this integral aspect of the data, with the USD-adjusted purchase amount representing the dependent variable.

The model incorporated a mix of numerical and categorical variables as predictors. After multiple refinements, the model ultimately utilized subscription duration, push notifications, auto renew, email subscriber, unique open count, demo user, open count, activity type count, ID count, purchase store, platform count, unique click count, and click count as predictive variables. A traditional train-test split method was employed, with the optimal results achieved using a 70/30 split. Overall, the model achieved an R^2 value of 0.64 and an MSE of 2,184.5. The individual R^2 values and coefficients for each predictor are displayed below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Coefficient** | **R^2** | **MSE** |
| **Overall Model** | **N/A** | **0.643** | **2141.82** |
| Subscription Duration | 0.005 | 0.610 | 2343.75 |
| Push Notifications\_No | 32.027 | 0.054 | 5680.11 |
| Push Notifications\_Yes | -32.027 | 0.054 | 5680.11 |
| Auto Renew\_Off | 39.254 | 0.050 | 5706.03 |
| Auto Renew\_On | -39.254 | 0.050 | 5706.03 |
| Email Subscriber\_No | 33.512 | 0.049 | 5710.74 |
| Email Subscriber\_Yes | -33.512 | 0.049 | 5710.74 |
| Unique Open Count | 10.385 | 0.010 | 5946.91 |
| Demo User\_No | 12.001 | 0.008 | 5955.53 |
| Demo User\_Yes | -12.001 | 0.008 | 5955.53 |
| Open Count | 5.181 | 0.008 | 5959.36 |
| Activity\_Type\_Count | -2.521 | 0.006 | 5967.14 |
| ID\_Count | -0.175 | 0.001 | 5997.56 |
| Purchase Store\_App | -42.376 | 0.001 | 5999.59 |
| Purchase Store\_Web | 42.376 | 0.001 | 5999.59 |
| Platform\_Count | -1.148 | 0.001 | 6002.29 |
| Unique Click Count | 0 | 0.000 | 6007.31 |
| Click Count | 0 | 0.000 | 6007.31 |

With an overall R^2 above 0.6, the model initially appeared promising for robust analysis. However, a closer examination of each predictor's individual R^2 values and coefficients revealed that more complex modeling would be required for definitive results. From an R^2 perspective, the only significant variable was subscription duration which exhibited an R^2 value above 0.6. In contrast, the remaining variables demonstrated negligible R^2 values.

Although the coefficient for subscription duration is relatively low in comparison to other variables, its high R^2 value indicates a positive linear relationship with the USD-adjusted purchase amount. This suggests that subscription duration is a critical factor in deriving conclusive insights.

**K-Means Clustering:**

K-Means is an unsupervised machine learning algorithm that operates by iteratively assigning each data point to one of K clusters based on the nearest mean, which serves as the cluster's centroid. Initially, these centroids are chosen randomly, and the algorithm recalculates them iteratively through a process of assigning data points to the nearest centroid and then updating the centroid to the mean of the newly assigned points. This process repeats until the centroids stabilize, indicating that the clusters are as internally homogeneous and externally distinct as possible given the value of K-clusters. To leverage this algorithm, the selection of K-clusters involved both domain knowledge and the Elbow Method, which looks for the point where increasing the number of clusters leads to diminishing returns in terms of within-cluster variance.

Given Rosetta Stone’s sparse dataset, K-Means clustering proved to be powerful and versatile method in identifying attractive customers to market additional products and services to. This unsupervised machine learning algorithm was used to partition the dataset into a set of distinct, non-overlapping subgroups, known as clusters, based upon different features from the dataset. Utilizing ‘Email Open %’, ‘Auto Renew’, ‘Subscription Duration’, and Purchase Amount’ – the k-means algorithm output, coupled with domain knowledge enabled in depth analysis in attractive customer segments.

**Random Forest Regression:**

Given the sparse data, yet numerous predictors, in Rosetta Stone’s dataset, a Random Forest Regression model was an additional tool leveraged to understand the data. This model utilizes ensemble learning, where it combines the output of multiple regression trees to produce a final result. The core idea behind this approach is that a group of weak learners (individual regression trees), when combined, can form a strong learner. Each tree in a Random Forest is built from bootstrap samples are drawn with replacement from the training set. This is done to ensure that each tree within ‘forest’ is comprised of different features, which encourages model robustness. Furthermore, the best split is chosen from the random subset of the features, ensuring that the most powerful predictor gets placed ‘higher’ in the tree. This strategy of combining random feature selection with bootstrapping helps in reducing the model's variance, making Random Forests a much more powerful tool for regression tasks compared to non-aggregative approaches.

This model implementation was necessary in understanding and identifying Rosetta Stone’s most valuable subscribers. While this model aims to predict the same continuous variable from the linear regression above, the innate differences in model computation allow for more nuanced data relationships to be introduced. There are two iterations present of this model, one for predicting ‘Subscriber Duration’, and an additional one for predicting ‘Purchase Amount’. Both of these target variables are key indications of a customer’s value to Rosetta Stone.

**Random Forest Classification Model:**

Random forest classifier operates under the same principles of ensemble learning as random forest regression; however, the main difference here lies in the type of dependent variable. Random forest classifiers predict a categorical variable, while the random forest regression predicts continuous variables. Each tree in the forest is built from the same bootstrapped sample drawn from the training set. This bootstrap aggregating approach combined with random feature selection helps achieve the same reduction in the variance of the model, making the Random Forest robust against overfitting – a common challenge in decision tree models. Its power in handling categorical variables comes from its ability to handle non-linear relationships and interactions between features without the need for explicit feature engineering, making it a versatile tool in diverse business applications where predictive accuracy and model interpretability are paramount. This robustness, coupled with its inherent ability to rank the importance of features, makes Random Forest an invaluable tool in analyzing complex relationships in the data and gaining meaningful insights for Rosetta Stone team.

This model implementation was essential in predicting customer churn based on a certain subset of features from the dataset. The model was built to predict churn based on ‘Subscription Duration’, ‘Purchase Amount’, ‘ID\_Count’, ‘Email Open %’, ‘Activity\_Type\_Count’, ’Demo User’, ‘Free Trial User’, & ‘Click\_Count’.

|  |  |  |
| --- | --- | --- |
| **Group** | **Subscriber Duration** | **USD-Adj. Purchase Amount** |
| 1 | Below 50th Percentile | Below 50th Percentile |
| 2 | Below 50th Percentile | Above 50th Percentile |
| 3 | Below 50th Percentile | Above 50th Percentile |
| 4 | Above 50th Percentile | Above 50th Percentile |

**Percentile-Based Segmentation:**

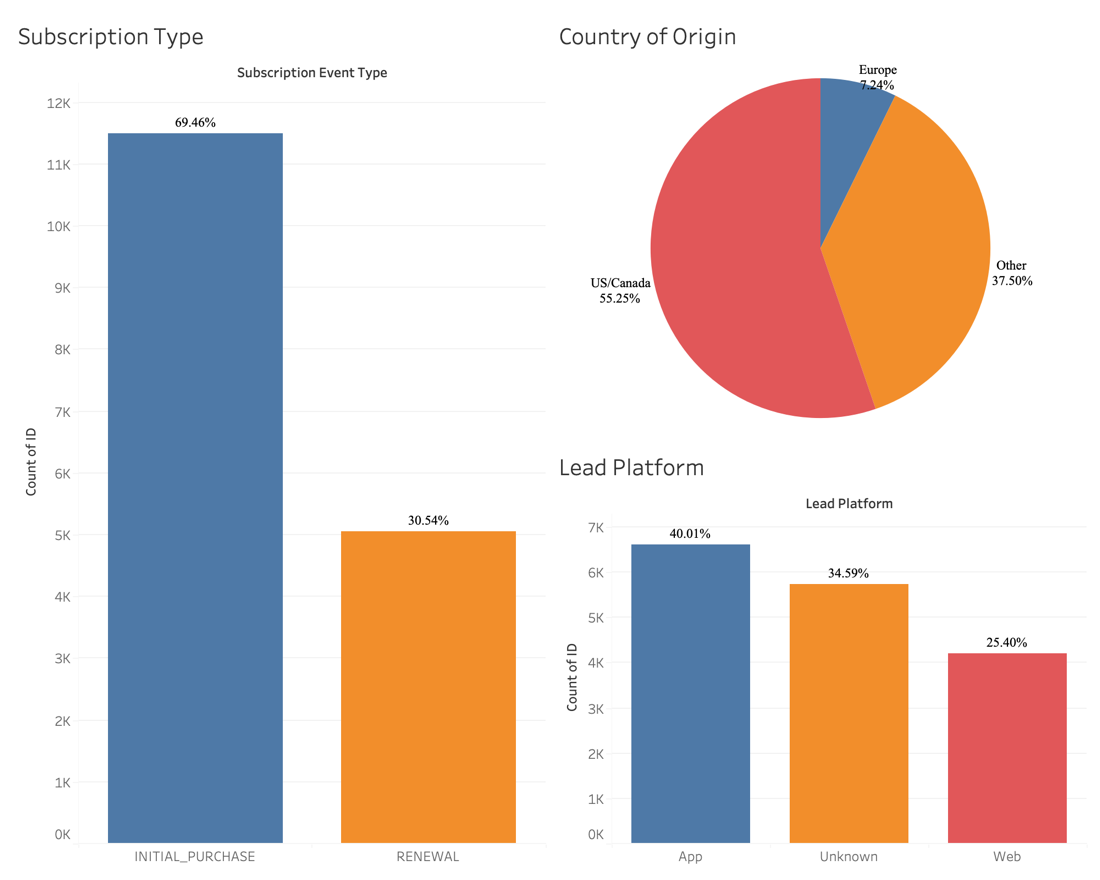
To address the limitations of linear regression and more complex models in discerning data segmentations, a non-modeling-based approach was adopted for segment identification. The contextual analysis of the data frame, informed by prior analyses such as linear and random forest regression, highlighted purchase amount and subscription duration as crucial variables for identifying valuable subscribers. Given Rosetta Stone's objective of generating returns, these findings align logically: a higher purchase amount indicates greater per-subscriber returns, and longer subscription durations suggest increased cumulative returns over time. This hypothesis was corroborated by the linear regression, which demonstrated a positive linear relationship between duration and purchase amount. Consequently, a segmentation strategy was developed, focusing on these two key criteria.

Using the `=quartile.exec()` function in MS Excel, the 50th percentiles for subscriber duration and USD-adjusted purchase amount were calculated as 39 and 365, respectively. Subscribers were then categorized into one of four segments based on their duration and purchase amount relative to these percentiles. The first segment included subscribers with both duration and purchase amount below the 50th percentile. The second segment comprised subscribers with duration above but purchase amount below the 50th percentile. The third segment captured those with duration below but purchase amount above the 50th percentile. Finally, the fourth segment consisted of subscribers who scored above the 50th percentile in both duration and purchase amount. The following breakdown summarizes these criteria and illustrates the distribution across these segments:

The segment distribution shows that groups 1 and 4 have a higher concentration of data points. This suggests a skewness in the data, leaning towards the extremes of low duration with low purchase amounts and high duration with high purchase amounts. Such a distribution further corroborates the positive linear relationship between duration and purchase amount. After establishing these groups, other variables were analyzed within each segmented group to identify key trends in the data, aiding in achieving the project objectives. Particularly, with subscribers in group 4 exhibiting both duration and purchase amounts above the 50th percentile, they emerged as the most valuable segment. The subsequent step involved pinpointing specific variables where group 4 distinctly stood out. This analysis is elaborated in the "Identifying Most Valuable Subscribers" section of the report.

**Subscriber Segments Dashboard:**

To gain more well-rounded understanding of the data, a visualization dashboard was created to summarize some of the key points. Within the Tableau dashboard, the platform type distribution was visually depicted using a bar chart, offering a quick overview of iOS, Android, and Web users. The addition of percentages enhanced the professional presentation and clarity of the insights, guiding decisions on user experience tailoring and marketing strategies. Another sheet focused on the geographic distribution of subscribers, utilizing color-coded pie chart to represent the percentage by country. Similarly, subscription type segmentation was addressed through visualizations showcasing the distribution of subscribers across different plans. By overlaying percentages, Tableau facilitated data-driven optimization for pricing and packaging decisions, providing stakeholders with a clear and professional presentation of insights.



**Addressing Objectives:**

**Understanding Subscriber Segments:**

In delving into the dataset, distinct subscriber segments have emerged, shedding light on essential patterns and trends that offer valuable insights into user behavior. One noteworthy observation is the discrepancy between initial subscriptions and renewal purchases. Initial subscriptions exhibit a higher frequency compared to renewals, suggesting a potential area for targeted strategies to boost renewal rates and foster long-term subscriber engagement. Geographical segmentation has revealed key focal points, with the US/Canada region emerging as the primary location for Rosetta Stone subscribers. Following closely are subscribers from the 'Other' category, indicating a diverse yet significant user base. Europe also stands out as a notable region, contributing to the global footprint of Rosetta Stone. Further segmentation based on lead platforms showcases distinctive user preferences. The app proves to be the most popular lead platform, indicating a strong inclination towards mobile language learning experiences. However, the presence of an 'Unknown' category signals an opportunity for exploration and understanding user behavior on platforms that are yet to be precisely identified. The website, while trailing behind the app, still plays a significant role in attracting and retaining subscribers. These insights into subscriber segments pave the way for strategic decision-making, allowing Rosetta Stone to tailor its offerings, marketing strategies, and user experiences based on the distinct preferences and behaviors observed within each identified segment. The identification of these trends provided crucial background knowledge for discerning model- and analysis-based segmentations.

The percentile-based segmentation, which categorized the data based on subscription duration and purchase amount, effectively highlighted four distinct segments within the subscriber database. These groups are outlined as follows:

**Segment 1 – Undesired Subscribers:** Positioned below the 50th percentile in terms of both subscription duration and purchase amount, these subscribers are deemed less valuable. Their tendency for shorter engagement and likelihood to cancel subscriptions, combined with their below-average returns, categorize them as undesirable.

**Segment 2 – Casual Subscribers:** Ranking above the 50th percentile for subscription duration yet below for purchase amount, these subscribers generally choose the most economical options available on the platform and tend to maintain their accounts for extended periods.

**Segment 3 – Cash Flow Subscribers:** Falling below the 50th percentile in subscription duration but above in purchase amount, these subscribers tend to spend more in a shorter timeframe on Rosetta Stone’s platforms. Although they contribute to immediate cash flow, their short, expected tenure may not support the long-term stability of user growth, with a probable higher turnover rate in memberships.

**Segment 4 – Desired Subscribers:** Exceeding the 50th percentile in both subscription duration and purchase amount, these customers are highly sought after by Rosetta Stone. They guarantee consistent and maximum returns, marking them as the ideal customer profile. The ensuing analysis and identification process for these subscribers are elaborated on in the subsequent section.

**Identifying Most Valuable Subscribers:**

As a business, Rosetta Stone's primary goal is to maximize the returns from each customer. When contextualizing the provided data, it is apparent that the subscribers contributing the highest returns are the most valuable. The purchase amount from each customer quantifies their individual return to Rosetta Stone. Recognizing this, the importance of this variable becomes clear in meeting the objective effectively.

Contextual analysis of the data revealed that a subscriber's duration is another significant indicator of value. Intuitively, subscribers with longer durations are expected to yield higher purchase amounts as they renew their subscriptions over time. This notion was empirically validated through modeling: a linear regression using subscriber duration as a predictor of USD-adjusted purchase amount showed a positive coefficient for this variable as well as exhibited the highest R^2 among the predictors. Consequently, it is evident that subscribers with both high purchase amounts and long durations are the most valuable. These subscribers typically fall into group 4 according to the percentile-based segmentation. The next phase involved exploring where group 4 distinguishes itself within the dataset. The analysis identified several factors that marked the distinct presence of group 4 as described below.

**Opting for Autorenewal:** As shown on the chart below, group 4 has the largest percentage of subscribers within its group that selected the autorenewal option.

Group 4 stands out with 76% of its subscribers choosing autorenewal, the highest proportion among all segments. This preference for autorenewal aligns with group 4's characteristics, as it suggests subscribers either make a deliberate choice to continue their Rosetta Stone experience after a significant period, or they may inadvertently maintain their subscription, which is a common occurrence in subscription-based services. Regardless of the reason, Rosetta Stone benefits from either scenario, as both contribute to longer subscriber durations and higher purchase amounts per customer. This decisively indicates that subscribers who opt for autorenewal are more valuable than those who do not.

**Having a High ID Count:** A notable variable derived from the app activity dataset and incorporated into the final dataset is the 'ID count.' This variable tallies the number of times a subscriber's ID appears in the dataset, revealing the frequency of their engagement with Rosetta Stone’s learning services. The chart below presents the average ID count for each percentile group, offering insights into usage patterns across different segments.

Group 4 sets itself apart with an average ID count of 7.4, the highest among all groups. This suggests that subscribers with a higher average ID count are likely to have longer durations and larger purchase amounts. Intuitively, this aligns with the expectation that subscribers who retain the platform for extended periods tend to interact with it more frequently, and those who invest more financially are inclined to maximize their usage. This conclusively demonstrates that subscribers who engage more actively with the platform are of greater value to Rosetta Stone.

**Being a Lifetime Subscriber:** The following chart breaks down the number of subscribers that fall into limited and lifetime categories segmented by their percentile-based grouping.

Group 4 stands out as the sole segment with a substantial proportion of lifetime subscribers, accounting for just over 30% within this category. In contrast, group 2, characterized by subscriber durations above the 50th percentile but purchase amounts below, is the only other segment with lifetime subscribers. This suggests that although a segment of lifetime subscribers may not generate high purchase amounts, a significant majority—approximately 75%—are classified within the optimal group 4. These findings compellingly suggest that lifetime subscribers are more likely to be of greater value.

**Further Insights on Subscriber Value from Random Forest Regression:** The random forest regression model yielded additional insights into the most significant factors contributing to high subscriber duration and purchase amounts. Two separate models were executed: the first aimed at predicting purchase amount, while the second focused on subscriber duration. The models' outputs highlighted the importance of each feature in predicting the respective outcomes. The feature importance results for these models are detailed below.

**A graph with blue bars

Description automatically generated**

**A graph with blue bars

Description automatically generated**

Group 4 sets itself apart with an average ID count of 7.4, the highest among all groups. This suggests that subscribers with a higher average ID count are likely to have longer durations and larger purchase amounts. Intuitively, this aligns with the expectation that subscribers who retain the platform for extended periods tend to interact with it more frequently, and those who invest more financially are inclined to maximize their usage. This conclusively demonstrates that subscribers who engage more actively with the platform are of greater value to Rosetta Stone.

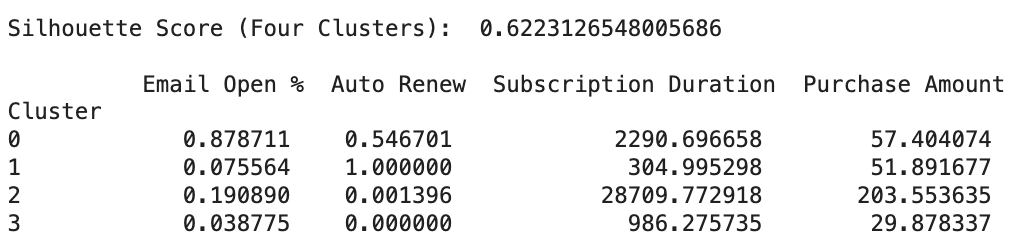
The random forest regression models highlighted 'email open percentage' as a key variable. This metric is computed by dividing each subscriber's open count by their send count, indicating the frequency at which a subscriber opens received emails. As depicted above, this variable emerged as the most influential factor in the model predicting purchase amount and the second most significant in predicting subscription duration. However, upon examining this variable through the lens of percentile-based segmentation, it did not significantly differentiate group 4 from the others, as can be seen in the following visualization.

Based on the random forest analysis, it was anticipated that group 4 would exhibit the highest average email open percentage among all groups. However, the actual results show that groups 2-4 have relatively similar values for this metric. This highlights certain limitations inherent in the percentile-based segmentation, suggesting that this analytical method might not always produce extremely precise outcomes. Therefore, more conclusive results can be obtained by integrating multiple analytical approaches. The fact that the percentile-based method and random forest regression generally align in many aspects convincingly validates the findings derived from these analytical techniques.

In summary, the most valuable subscribers for Rosetta Stone are those with longer subscription durations and higher purchase amounts. These subscribers typically display specific characteristics: they opt in for autorenewal, exhibit high ID counts, fall into the lifetime subscriber category, and demonstrate a high email open percentage.

**Identifying Subscribers with Upselling Potential:**

As aforementioned, the sparsity of the dataset makes it difficult to leverage supervised models to segment customers who may have a propensity to purchase additional goods and services from Rosetta Stone. To expand on our prescribed customer segments, the K-Means model was built to obtain a more granular view of the customer base and the hidden patterns present. The model abstracted and implemented utilized the features, ‘Email Open %’, ‘Auto Renew’, ‘Subscription Duration’, & ‘Purchase Amount’. While these are four distinct predictors, each of them was addressed above as holding high importance in predicting ‘Purchase Amount’ & ‘Subscription Duration’ – and holds impressive value in further segmenting the customers based on their engagement with Rosetta Stone’s marketing efforts and prolonged use of the platform. After tuning the hyperparameters of the model using within cluster sum of squares plot (elbow method), it was determined that the optimal number of clusters was four, and the average feature value of each cluster is present below.



In the context of K-Means clustering, the silhouette score is a measure of how similar the customer’s four predictor values is to its own cluster compared to external clusters. The value ranges from -1 to 1, where a high value indicates that the customer is well matched to its own cluster and poorly matched to neighboring clusters. A silhouette score of 0.62 suggests relatively strong cluster cohesion and separation – meaning that – on average, each customer in each cluster is more similar to their own cluster than to the customers in other clusters. This implies that the K-Means model has done a reasonably good job at identifying distinct and relevant groups of customer segments within the data.

Now that the K-Means algorithm’s clustering capabilities are proven, there are some interesting patterns present within each customer segment:

Cluster 0 – The Loyal Subscribers

* Cluster 0 is comprised of Rosetta Stone’s most attractive customer segment to market additional offerings to. This customer base has the highest percentage of marketing email engagement (~.88), an exceptional number of users on auto renew (~.55), long subscription duration (2290 days), and a fair average purchase amount ($57.40).
* Because of how likely a customer in this segment is to both engage with Rosetta Stone’s outreach and have auto renew turned on – this cluster becomes extremely attractive and Rosetta Stone executives should aggressively market new products and offerings to this customer segment.

Cluster 1 – The ‘Free-Trial’ Forgetters

* Cluster 1 has a slightly less-attractive prospective customer base than cluster 0. This customer base does not have a very high propensity to open emails (~.08), perfect auto renew engagement (1.0), low subscription duration comparatively to other clusters (305 days), yet still a fair average purchase amount ($51.89).
* This customer base that doesn’t seem to engaged with their subscriptions. While this customer base has a ‘perfect’ auto renew percentage, these subscribers do not open Rosetta Stone marketing emails and have the lowest subscription duration across the four clusters.
* This cluster resembles likely values across the four features of a customer who signed up for a free trial and likely forgot to cancel their subscription prior to the culmination of their free trial. Rosetta Stone executives and marketers should neglect offering new products or services to this customer segment.

Cluster 2 – The Lifetime Subscribers

* Cluster 2 is comprised of almost entirely lifetime subscribers. Members of this cluster have fairly high email engagement (~.19), a very low percentage of customers with auto renew (~.0014), have an average subscription duration of (28710 days), and the highest purchase amount ($203.55).
* The culmination of these statistics proves to be the second customer base that Rosetta Stone should target. Although they may be less likely to purchase products in line with current offerings (since this customer base purchased a lifetime membership), this customer base’s encouraging email engagement and lifetime loyalty to the brand creates a reasonable probability for success selling additional products.

Cluster 3 – The ‘Noisy’ Customer Base

* Whereas the previous three customer segments could be identified from collating domain/contextual business knowledge with clustering results, this cluster seems quite noisy.
* This cluster has an impressively low engagement with Rosetta Stone’s marketing efforts (~..03), not a single customer has auto renew turned on, yet they have a large subscription duration (986 days), and low purchase amount (~$29.88).
* Because this cluster’s aggregated statistics fails to belong to a holistic customer segment, any additional customers that get placed into this cluster should not be targeted for additional products and services.

With the above discussion, it was apparent that two attractive customer bases come to fruition. Cluster 0 (‘The Loyal Subscribers’) and Cluster 2 (‘The Lifetime Subscribers’) should be heavily targeted by Rosetta Stone marketers to sell additional products to.

**Indications of Subscriber Churn and Potential Solutions:**

*Known Limitations & Assumption*: Present in the dataset, is the categorical variable, ‘Auto Renew.’ While there isn’t a perfect metric in this dataset for tracking customer churn – customers who do not have auto renew on will automatically be churned at the end of their billing cycle. To predict customer churn and understand the important features in explaining the variance in the target variable – a logistic regression model and random forest classifier predicting auto-renew was implemented.

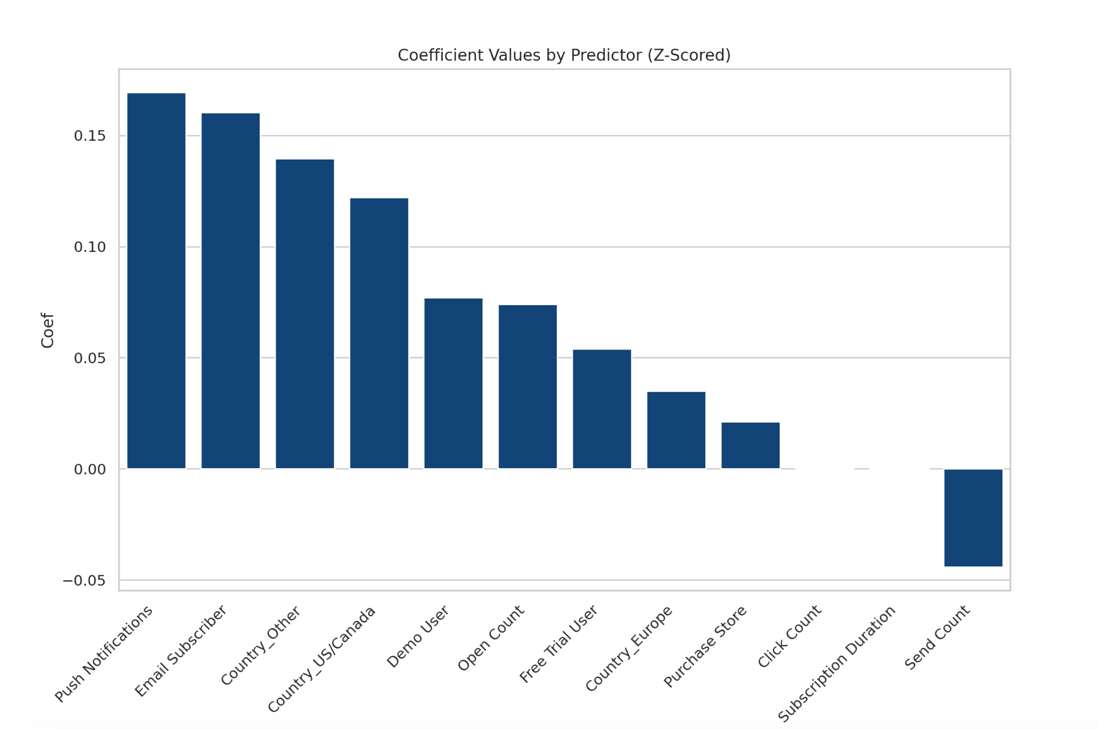
**Logistic Regression**

Through the logistical regression analysis, it is possible to identify certain subscriber profiles that are less likely to continue usage with Rosetta Stone, as well as pinpoint potential barriers to deeper subscriber engagement where possible.

Subscribers who exhibit a lower likelihood of auto-renewal, as indicated by the negative coefficients in the model, often share common characteristics. Various variables were used, including ‘Purchase Store’, ‘Demo User’, ‘Free Trial User’, ‘Demo Subscriber’, ‘Send Count’, Subscription Duration’, etc. By analyzing the odds coefficients, we can determine whether those variables had a positive or negative impact on the ‘Auto Renew’ variable. For instance, the negative coefficient associated with 'Send Count' suggests that subscribers receiving an excessive number of notifications or emails might feel overwhelmed or experience communication fatigue, leading them to disengage from the service. This finding points to a critical barrier in subscriber engagement—over-communication—which can be perceived as intrusive or spammy, thereby discouraging continued use.

Additionally, the slight negative influence of longer 'Subscription Duration' on renewal likelihood hints at another potential barrier: the need for reevaluation over time. Subscribers on longer plans might be more inclined to reassess the value and relevance of the service as their subscription progresses, potentially leading to disengagement if their evolving needs are not being met.

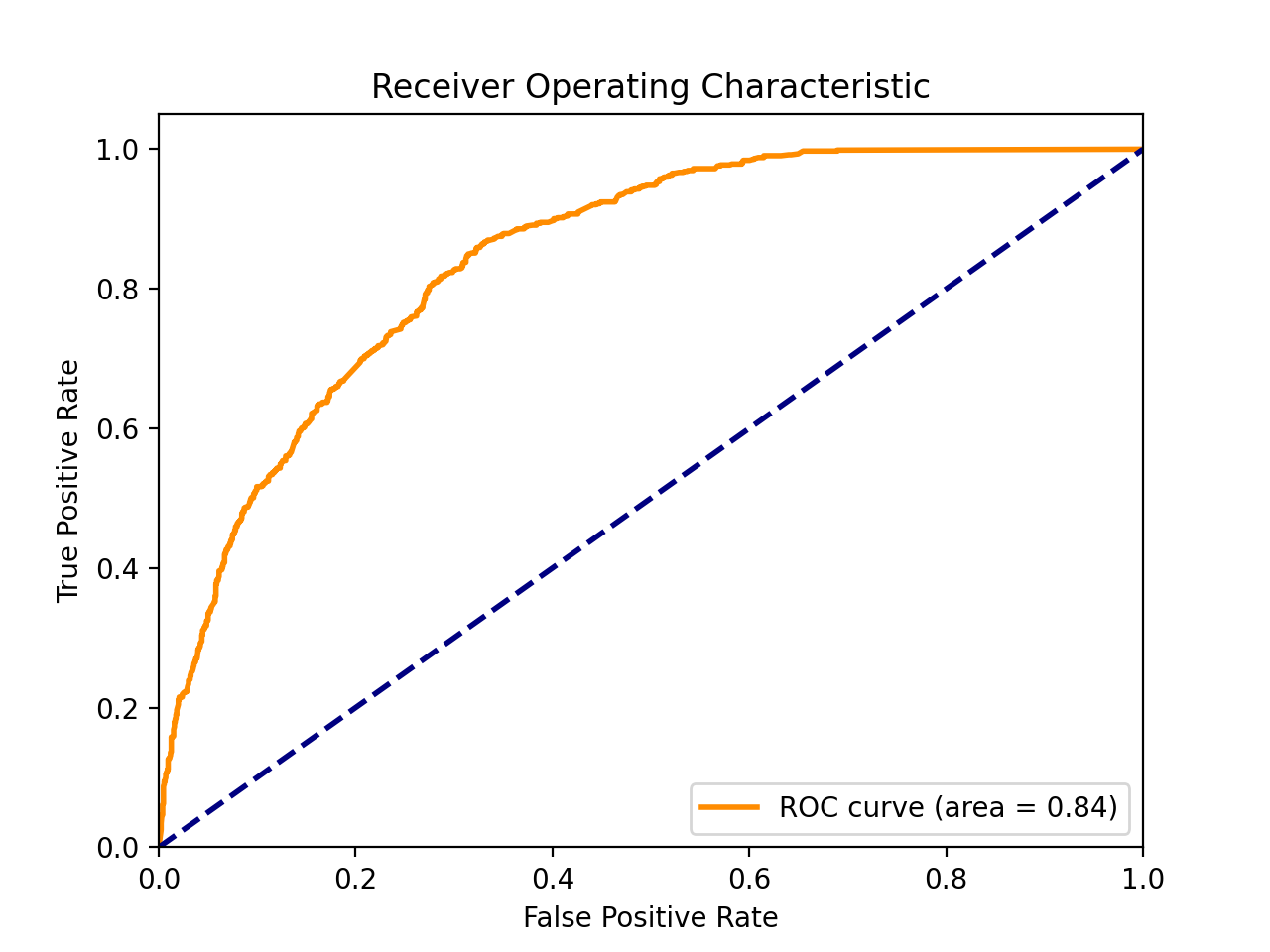
These insights imply that for certain subscriber groups, particularly those inundated with communications or those on extended subscription plans, there are clear areas where engagement strategies could be refined. Addressing these barriers—by optimizing communication frequency and ensuring ongoing service relevance and value—could enhance overall subscriber satisfaction and retention.



**Random Forest Classification**

While the logistic regression model gives us some powerful insight into which variables attribute positive or negative likelihood of auto renew being on (thus a customer churning). The random forest classification model gives us additional insight into each predictor’s generalized importance in determining churn, while additionally enabling predictive power in new customers outside of the instances (customers) present in this dataset.

While it was mentioned above in the report, this model was trained built to predict churn based on ‘Subscription Duration’, ‘Purchase Amount’, ‘ID\_Count’, ‘Email Open %’, ‘Activity\_Type\_Count’, ’Demo User’, ‘Free Trial User’, & ‘Click\_Count’. The model is comprised of five hundred individual decision trees, which when aggregating each tree’s feature importance, we gain tremendous ability to predict customer churn based.



The figure above is a Receiver Operating Characteristics plot (typically called ROC/AUC) – which is a direct visualization of the model’s predictive capabilities. The model correctly classifies churning customers with 84% accuracy – meaning that marketers, analysts can run a future customer through this model and correctly predict if the customer will churn at the end of their billing cycle 84% of the time. This model additionally provides further insight into which features may indicate a customer’s propensity to churn.

A bar graph with blue bars

Description automatically generated

The feature importance bar plot provided further insight into which predictors are most causal to customer churn. The top four important features although slightly nuanced, all hold a relation to engagement on Rosetta Stone’s platform. While these features do denote the overall ‘importance’ in this model’s classification, the logistic regression model above provides additional insight as to how the features impact likelihood of negative or positive classification.

**Additional Business Opportunities:**

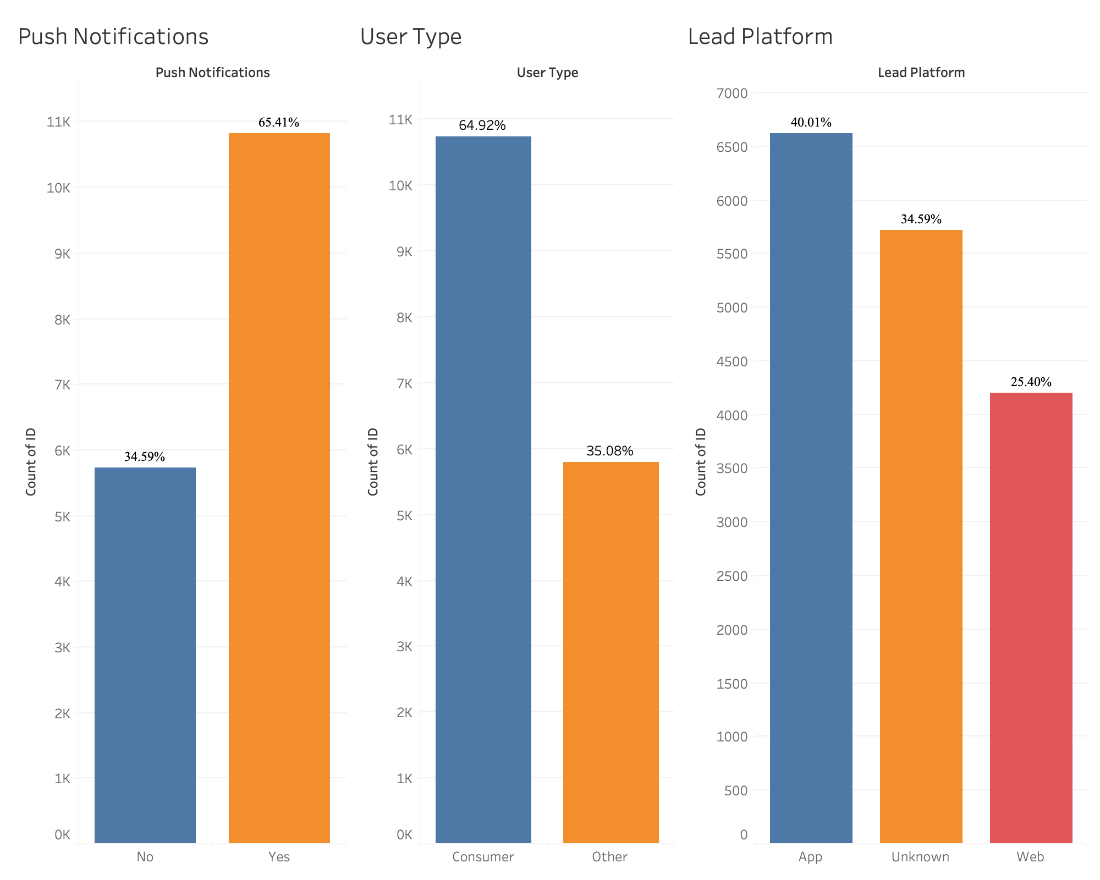
In the pursuit of identifying strategic avenues for growth, delving into the complex interplay of variables and their impact on ID Count within the dataset. Core elements, such as lead platforms, user types, and push notifications, took center stage in the exploration. Through various visualizations, we dissected the influence of different lead platforms on ID Count, discerned the contributions of various user categories, and quantified the impact of push notifications on user engagement.

This focused analysis served as the basis for practical recommendations. Insights gleaned from ID Count informed strategic refinements, optimizing communication channels, tailoring user experiences, and streamlining engagement initiatives. By understanding the intricate dynamics between push notifications, user categories, and lead platforms, our strategies were precisely calibrated to align with user preferences, fostering satisfaction and retention.

The idea of leveraging subscriber data and monetizing it presents a multitude of prospects. An in-depth analysis of subscriber data offers a valuable opportunity for monetization and strategic utilization. By delving into language learning trends, consumer behaviors, and emerging patterns, insightful reports can be curated and leveraged. These reports hold the potential to cater to diverse entities such as educational institutions, businesses, and industry competitors, offering them comprehensive insight(s) reports. In turn, this initiative has the potential to unlock additional supplementary revenue streams which provides the opportunity for reinvestment in continual overall growth.

Elevating the language learning experience through enhanced in-app gamification and fostering consumer-centric language learning communities emerges as a strategic imperative for Rosetta Stone. With a substantial user base actively engaging with push notifications, the opportunity to amplify user interaction by introducing gamification elements becomes apparent. Leveraging quests, badges, and challenges can not only make language learning more immersive and enjoyable but also capitalize on the heightened engagement driven by push notifications. These initiatives, designed to enhance user satisfaction and engagement, play a pivotal role in customer retention.

Moreover, tailoring strategies to cater to the larger consumer user base not only helps in retaining existing customers but also holds the potential to increase customer spending. For instance, incorporating features like purchasing extra lives for challenges through in-app transactions or offering premium badges for language mastery can incentivize users to spend more within the Rosetta Stone ecosystem. This dual benefit of retaining and monetizing existing customers positions these business opportunities as valuable contributors to the overall growth and profitability of Rosetta Stone. The integration of gamification elements not only enriches the language learning journey but also strategically encourages continued user engagement and financial investment in the platform.



**Finalized Appendix:**

1: **Data Cleaning and Merging**

1.1: Updated Data Dictionary

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Source |
| ID | Individual ID for Each Subscriber | Original Dataset |
| Language | language selected by subscriber | Original Dataset |
| Subscription Type | Limited = subscription has a defined expiration, usually 12 or 24 months from start date; Lifetime = perpetual subscription, indicated by end date in year 2098 or 2099 | Original Dataset |
| Subscription Event Type | Type of purchase; initial purchase or renewal | Original Dataset |
| Purchase Store | Web or app | Original Dataset |
| Purchase Amount | Amount paid for subscription, available for web purchases only | Original Dataset |
| Currency | Currency for purchase amount | Original Dataset |
| Subscription Start Date | Date subscription begins | Original Dataset |
| Subscription Expiration | Date subscription ends | Original Dataset |
| Demo User | Subscriber has used demo content on the app | Original Dataset |
| Free Trial User | Subscriber registered for a limited time free trial (usually 3 days) | Original Dataset |
| Free Trial Start Date | Date trial subscription begins | Original Dataset |
| Free Trial Expiration | Date trial subscription ends | Original Dataset |
| Auto Renew | Flag to indicate if subscriber has turned off the auto renew option | Original Dataset |
| Country | Country where subscriber lives (self reported) | Original Dataset |
| User Type | Consumer or other (homeschool, business) | Original Dataset |
| Lead Platform | Platform subscriber used to engage with Rosetta Stone products (web, app, or unknown) | Original Dataset |
| Email Subscriber | Subscriber is opted in to receive emails | Original Dataset |
| Push Notifications | Subscriber is opted in to receive push notifications | Original Dataset |
| Send Count | Number of emails sent to subscriber in past 90 days | Original Dataset |
| Open Count | Total number of times emails were opened by subscriber in past 90 days (of those sent during same time period) | Original Dataset |
| Click Count | Total number of times emails were clicked by subscriber in past 90 days (of those sent during same time period) | Original Dataset |
| Unique Open Count | Unique number emails opened by subscriber in past 90 days (of those sent during same time period) | Original Dataset |
| Unique Click Count | Unique number emails clicked by subscriber in past 90 days (of those sent during same time period) | Original Dataset |
| ID\_Count | ID\_Count – The number of times an ID is mentioned in app activity | Derived/Modified |
| USD\_Adjusted\_Amount | USD\_Adjusted\_Amount – purchase amount adjusted by currency exchange rates | Derived/Modified |
| Subscription Duration | Subscription Duration – the number of days from subscription start date to end date | Derived/Modified |
| Activity\_Type\_Count | Activity\_Type\_Count - Derived or modified during data cleaning and analysis | Derived/Modified |
| Language\_[Insert Language Code] | Language\_[Insert Language Code] - 1 for any instance with included language, 0 for different language | Derived/Modified |

1.2: Table of currency exchange rates which were utilized to convert purchase amount values to USD. Sourced from Wise.com on 11/10/2023.

|  |  |
| --- | --- |
| **Currency** | **Exchange Rate to USD** |
| GBP | 1.22285 |
| PLN | 0.24706 |
| EUR | 1.0723 |
| BRL | 0.20375 |
| KRW | 0.00077 |
| MXN | 0.05692 |
| UAH | 0.02765 |
| RUB | 0.01105 |
| CHF | 1.10693 |
| CAD | 0.72474 |
| CZK | 0.04434 |
| RON | 0.21426 |
| AUD | 0.6516 |
| THB | 0.0282 |
| JPY | 0.00659 |
| INR | 0.01202 |
| COP | 0.00025 |
| BGN | 0.54737 |
| ZAR | 0.05341 |
| HUF | 0.00281 |

**2: Modeling**

2.1: Linear Regression Code

from google.colab import drive

drive.mount('/content/drive')

file\_path = '/content/drive/MyDrive/Cleaned\_Subscriber\_Information (1).csv' # Adjust the path if your file is in a different folder within Drive.

# Now you can use pandas or any other library to read the file.

import pandas as pd

df = pd.read\_csv(file\_path)

# Data processing and manipulation

import pandas as pd

# Machine learning model and data splitting

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

# Model evaluation metrics

from sklearn.metrics import mean\_squared\_error, r2\_score

# Select the columns for linear regression analysis

columns\_to\_use = ['Subscription Type', 'Purchase Store', 'Demo User',

'Auto Renew', 'Email Subscriber', 'Push Notifications', 'Subscription Duration', 'USD\_Adjusted\_Amount',

'Open Count', 'Click Count','Unique Open Count','Unique Click Count','ID\_Count','Platform\_Count','Activity\_Type\_Count']

# Create a new DataFrame with the selected columns

df\_lr = df[columns\_to\_use].dropna()

# Create dummy variables for the categorical columns

df\_lr = pd.get\_dummies(df\_lr, drop\_first=False)

# Separate the independent variables (X) and the dependent variable (y)

X = df\_lr.drop('USD\_Adjusted\_Amount', axis=1)

y = df\_lr['USD\_Adjusted\_Amount']

# Split the data into training and testing sets (70/30 split)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

# Create and train the linear regression model

model = LinearRegression()

model.fit(X\_train, y\_train)

# Predict on the test set

y\_pred = model.predict(X\_test)

# Calculate R² and MSE for the model

r2 = r2\_score(y\_test, y\_pred)

mse = mean\_squared\_error(y\_test, y\_pred)

# Print the results

print(f'R² Score: {r2}')

print(f'Mean Squared Error (MSE): {mse}')

results\_df = pd.DataFrame(columns=['Variable', 'Coefficient', 'R2', 'MSE'])

# Training a model for each variable and storing the results

for col in X\_train.columns:

X\_train\_col = X\_train[[col]]

X\_test\_col = X\_test[[col]]

model = LinearRegression()

model.fit(X\_train\_col, y\_train)

y\_pred = model.predict(X\_test\_col)

coeff = model.coef\_[0]

r2 = r2\_score(y\_test, y\_pred)

mse = mean\_squared\_error(y\_test, y\_pred)

results\_df = results\_df.append({'Variable': col, 'Coefficient': coeff, 'R2': r2, 'MSE': mse}, ignore\_index=True)

# Displaying the results

results\_df.sort\_values(by='R2', ascending=False)

2.2: Percentile Based Segmentation

**Percentile Formula:**

=QUARTILE.EXC(Y$2:Y$10043,2)

**Application to Data Formula:** =IF(AND(Y2<=$AI$8,Z2<=$AJ$8),1,IF(AND(Y2<=$AI$8,Z2>$AJ$8),2,IF(AND(Y2>$AI$8,Z2<=$AJ$8),3,IF(AND(Y2>$AI$8,Z2>$AJ$8),4,0))))

2.3.1: Random Forest Regression – Predicting Subscription Duration

A screen shot of a computer

Description automatically generated

2.3.2: Random Forest Regression – Predicting Purchase Amount

A screen shot of a computer program

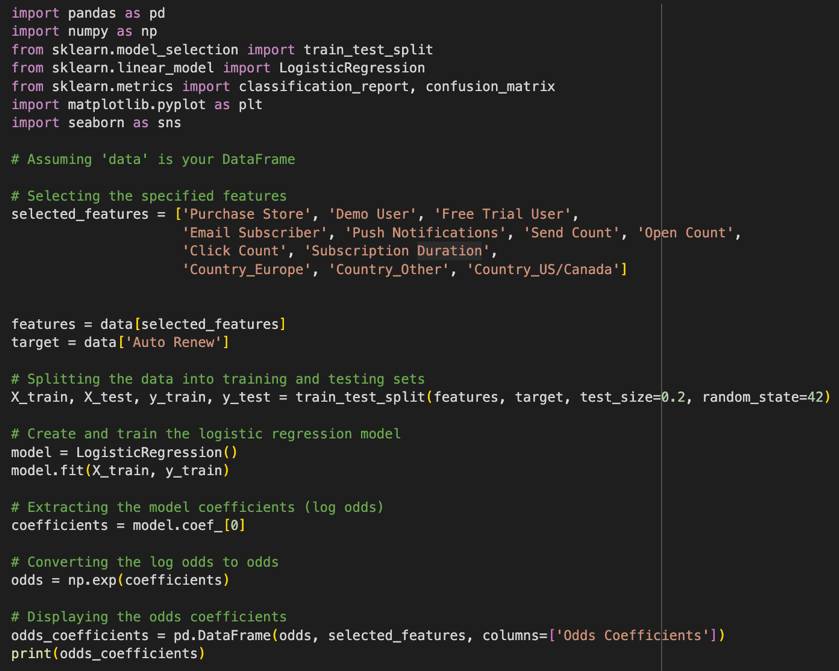
Description automatically generated

2.4: K-Means Hyperparameter Tuning, Final Model Creation, and Model Evaluation



2.5: Random Forest Classification Model – Predicting Customer ChurnA screenshot of a computer program

Description automatically generated2.6: Logistic Regression Code



3.1: Additional Visualizations

