



F.A.T.T.Y.

BRSM Team Project

Report By

Aryan Gupta - 2021113012

Team

Aryan Gupta - Bhvya Kothari - Ayush Yadav

Introduction

This analysis is centered on a dataset [1] obtained from Kaggle, which originates from a mobile application operated by a Fintech company. Their application offers a platform where users can manage their finances more effectively and opt into premium services for enhanced features. The primary objective of utilizing this dataset is to explore and identify patterns and behaviors among users that could drive increased engagement and revenue for the fintech company.

Objective

The project aims to test specific hypotheses using the dataset to better understand the consumer base of the fintech app. These hypotheses focus on determining whether there are significant differences in app usage behaviors between users who subscribe to premium services and those who do not.

Insights derived from this analysis may be used for targeted marketing strategies and product development to boost user engagement and increase revenue.

By identifying trends and patterns in user behavior and demographic characteristics, the company can tailor its offerings more effectively, thus enhancing user satisfaction and conversion rates.

The Data Set

The dataset provides a comprehensive view of user interactions within the app, comprising multiple attributes captured during user sessions. It features 50,000 entries, each representing an individual user, and includes a variety of data types:

- Binary Data: Indicating yes/no decisions, such as enrolled for premium service subscription and used_premium_feature to denote usage of premium capabilities.
- Numerical Data: Such as age of the users and num screens reflecting the count of screens accessed during a session.
- Categorical Data: Encoded numerically for ease of analysis, including demographic indicators like ProfileMaritalStatus and ProfileEducation.

Significant column:

- enrolled: Indicates whether a user has subscribed to premium services.

Methods

Hypothesis:

1. Is there a statistically significant difference in the mean age between users who enroll in the premium service and those who do not?
2. Do users who enroll in the premium service visit a significantly different number of screens compared to those who do not enroll?
3. Is there a significant association between the usage of premium features during the free trial and the likelihood of enrolling in the premium service?
4. Does the day of the week when a user first opens the app have a significant influence on their decision to enroll in the premium service?
5. Does engagement with mini-games within the app significantly affect the likelihood of users enrolling in the premium service?
6. Is there a significant correlation between users liking features in the app and their likelihood of enrolling in the premium service?

Work Distribution Among Teammates

H1 & H2 - Aryan Gupta

H3 & H5 - Ayush Yadav

H4 & H6 - Bhvya Kothari

Tools/Tests used to confirm or reject hypothesis:

1. Regression Tests
2. ANOVA
3. T-Test
4. Chi-Square Tests
5. Mann-Whitney U test
6. Correlation Matrix / Pearson Correlation Coefficient
7. Shapiro-Wilk test

1) Data Cleaning and Preprocessing

We began our study by carefully cleaning up the data. We checked different factors and how they were spread out in the customer data. This included fixing any missing information, identifying any unusual data points, and making sure all the data was in a consistent format. Our initial goal was to make sure the data we used for our analysis was accurate and reliable, which helps make our study results strong and trustworthy. Additionally, we conducted several normality tests and created distribution plots to determine which variables were most appropriate for our analysis. During this process, we identified certain columns, which exhibited uniform distributions throughout. Consequently, we made the decision to exclude these columns from further consideration, as they did not contribute significantly to our analytical objectives.

2) Correlation

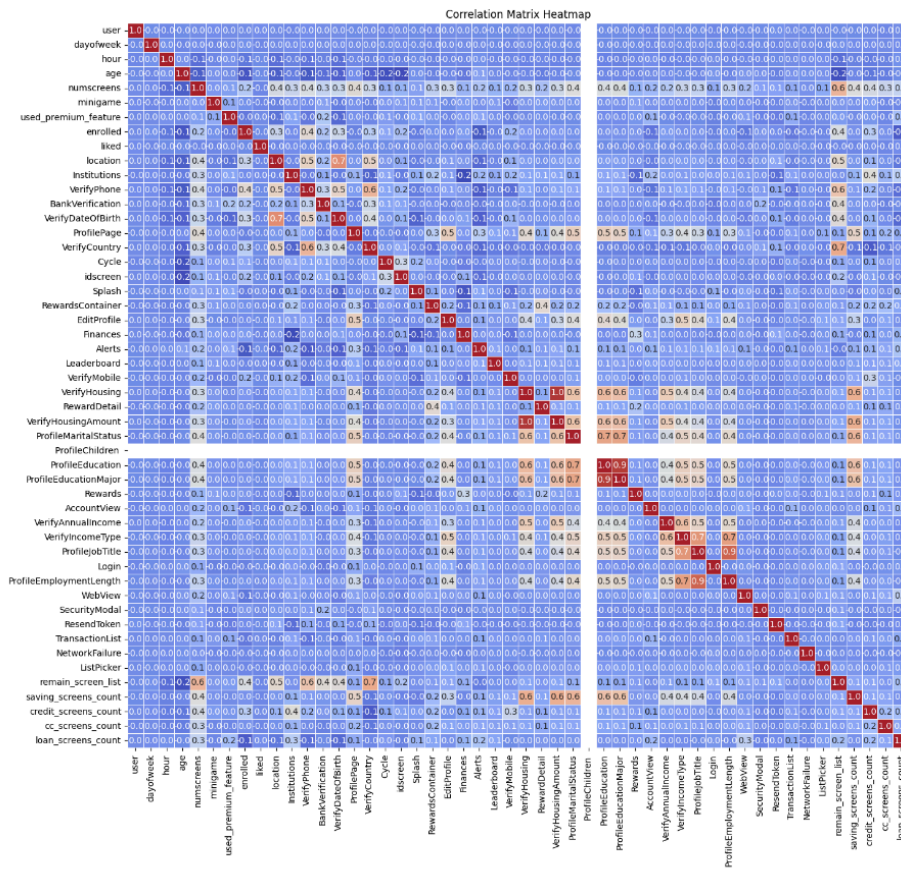


Fig. 1
Correlation
Heatmap

Collinearity denotes a scenario wherein two or more predictor variables within a regression model exhibit a high degree of correlation, potentially leading to statistical challenges such as the instability of estimates and inaccuracies in regression coefficients. In our study, we undertake an examination of the correlation heatmap utilizing the Pearson correlation coefficient to discern pairs of features characterized by elevated correlation coefficients, indicative of multicollinearity within the dataset. The heatmap, depicted in Figure 1, furnishes a graphical depiction of the interrelation among various feature groups encompassing all attributes across the dataset. Typically, the most pronounced collinearity exists among variables within the same feature group; however, instances of robust correlation between variables from disparate feature groups can also be identified.

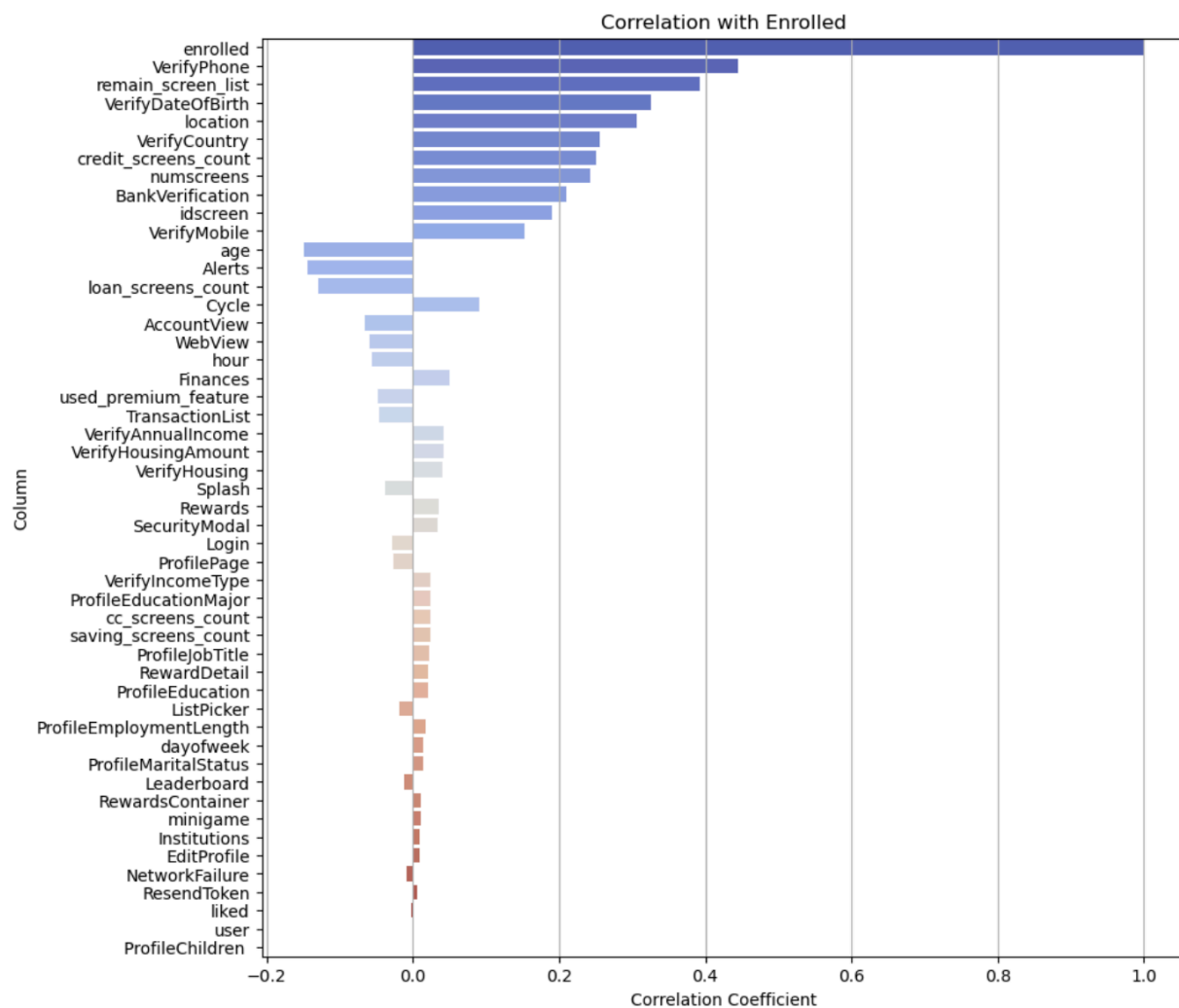


Fig 2. Correlation with Enrolled

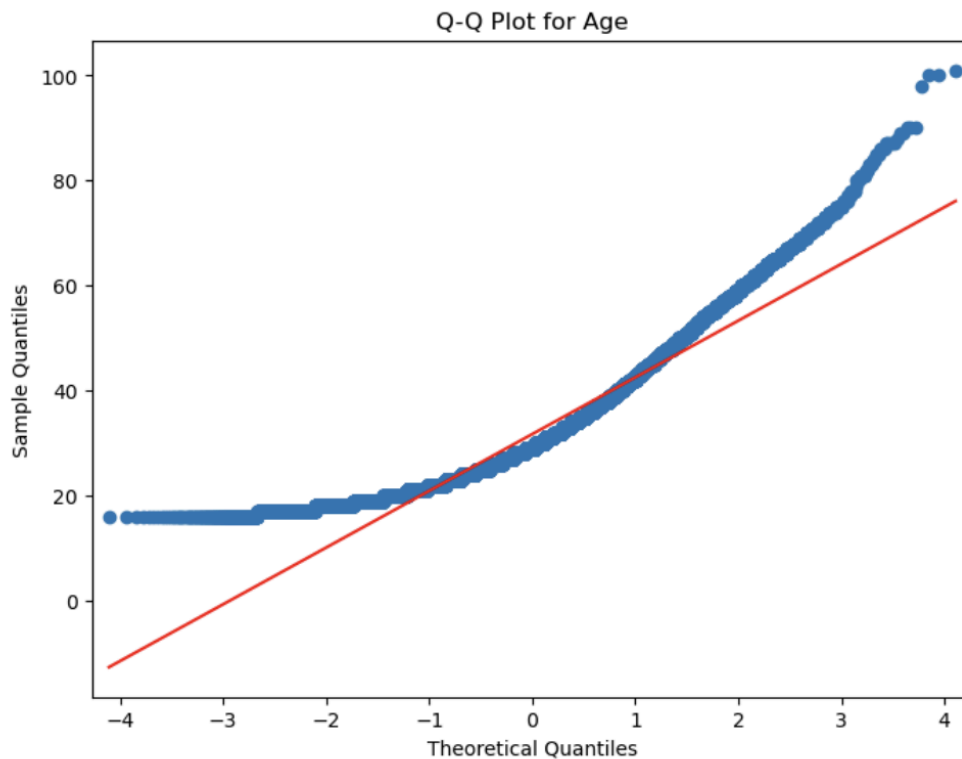
Figure 2 highlights variables which are highly correlated with our target attribute Enrolled, hence these columns of the dataset are more important to study. Many of the variables highly positively correlated with "enrolled" exhibit a binomial nature, such as provided

location, bank verification, and verify phone number. This observation aligns logically with the premise that individuals who are enrolled, are obliged to verify and furnish such personal information. Consequently, it is reasonable to infer that many non-enrolled users may have opted not to divulge their data, thereby contributing to this correlation pattern.

Hypothesis 1

Is there a statistically significant difference in the mean age between users who enroll in the premium service and those who do not?

In this part of the study we try to analyze how age affects enrollment status. Clearly by looking at the correlation matrix above, we do have the intuition that age negatively affects enrollment status, here we try to test and formulate it. The null hypothesis being the distribution of ages of users who are enrolled and who are not enrolled are the exact same.

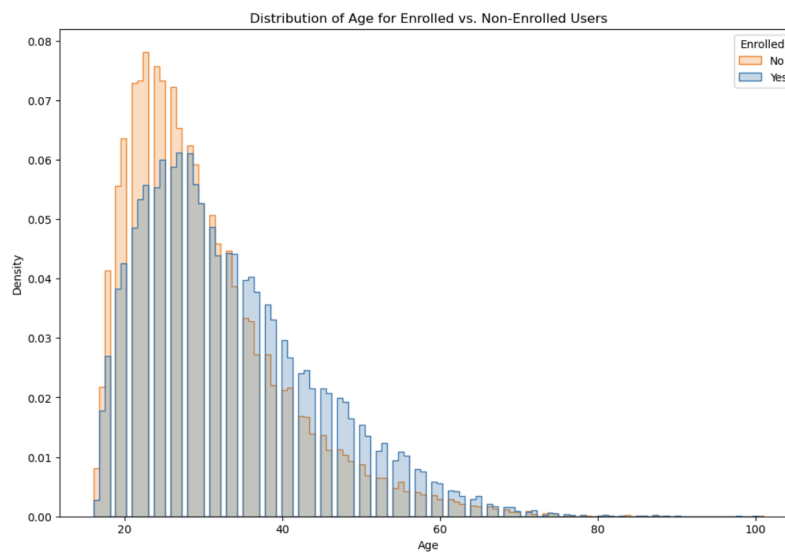


Shapiro-Wilk Test results:

```
ShapiroResult(statistic=0.9183380007743835, pvalue=0.0)
```

Fig 3. Q-Q plot & Shapiro test result

Based on Fig 3, We can infer that the age distribution does not follow a normal distribution, as the p-value of the shapiro test [2] is significantly below 0.05, rejecting the null hypothesis of normality. Given this non-normality of the age data, a non-parametric test (Mann-Whitney U test), which does not assume normal distribution, was used to compare the mean ranks of the ages between users who enroll in the premium service and those who do not.



Mann-Whitney U test result: MannwhitneyuResult(statistic=256683959.5, pvalue=2.3910540280338208e-262)

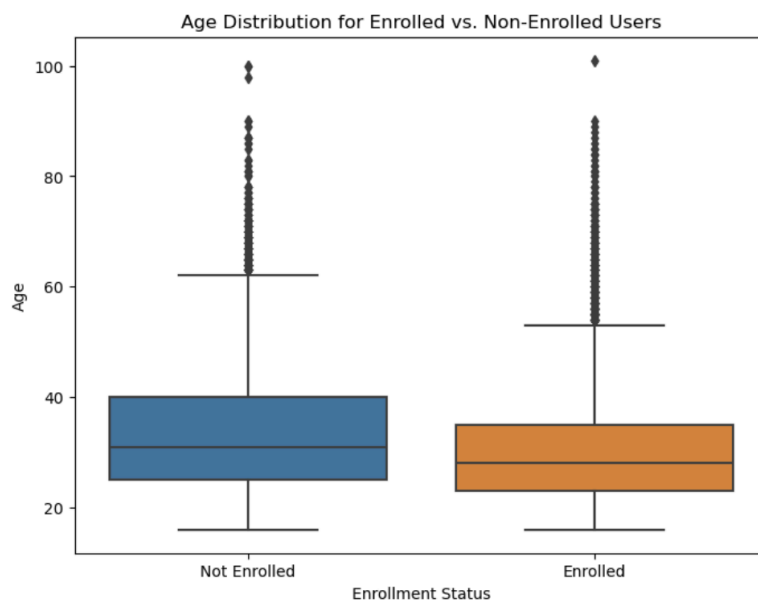


Fig 4. Histogram of distribution of age with respect to enrolled and non enrolled

Fig 5. Mann-Whitney U test result & boxplot

Mann-Whitney U Test [4] Result, statistic is 256683959.5 with a p-value of 10^{-262} . This extremely small p-value indicates a statistically significant difference in the age distributions between users who are enrolled and those who are not enrolled in the premium service. This supports the hypothesis that there is a difference in mean age between the two groups.

The boxplot shows that non-enrolled users tend to be older compared to enrolled users. The median age of non-enrolled users appears higher, and the age distribution is tighter among enrolled users. The histogram also supports the same.

Hence, **there is a statistically significant difference in the mean age between users who enroll in the premium service and those who do not.** Specifically, younger users are more likely to subscribe to the premium service compared to older users.

Hypothesis 2

Do users who enroll in the premium service visit a significantly different number of screens compared to those who do not enroll?

In this part of the study we try to analyze how the number of screens our user uses affects enrollment status. Clearly by looking at the correlation matrix above, we do have the intuition that the number of screens positively affects enrollment status, here we try to test and formulate it. The null hypothesis being the distribution of screens of users who are enrolled and who are not enrolled are the exact same.

Shapiro-Wilk Test results for 'numscreens': ShapiroResult(statistic=0.8670444488525391, pvalue=0.0)

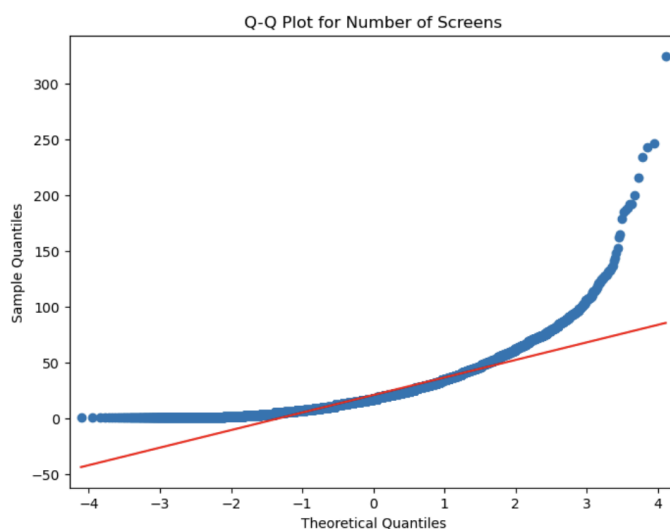
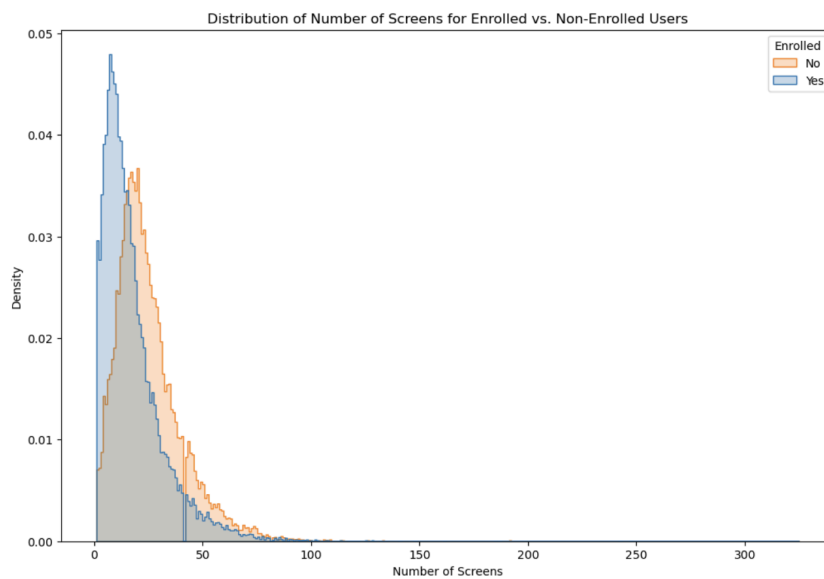


Fig 7. Q-Q plot and Shapiro-Wilk Test

Based on Fig 7. The Shapiro-Wilk test result for the number of screens (numscreens) shows a statistic of 0.867 and a p-value of 0.0, indicating that the distribution is not normal. The Q-Q plot confirms the results of the Shapiro-Wilk test, showing a clear deviation from the theoretical normal distribution line, especially in the tails. This deviation suggests that the data is skewed and not normal. Given this non-normality of the data, a non-parametric test (Mann-Whitney U test), which does not assume normal distribution, was used to compare the mean ranks of the ages between users who enroll in the premium service and those who do not.



Mann-Whitney U test result for Number of Screens: MannwhitneyuResult(statistic=422652951.0, pvalue=0.0)

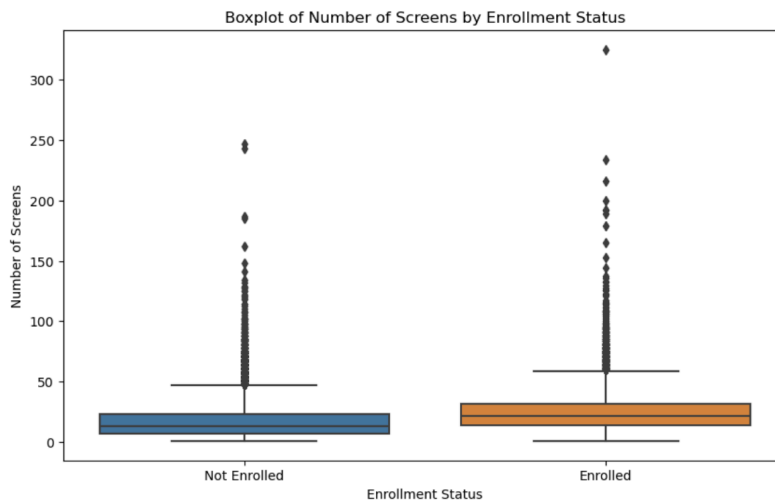



Fig 4. Histogram of distribution of Screens with respect to enrolled & non-enrolled

Fig 5. Mann-Whitney U test result & boxplot



From figure 4, The histogram for the number of screens visited shows a clear difference in the distribution between enrolled and non-enrolled users. Enrolled users tend to visit more screens compared to non-enrolled users, as indicated by the right shift of the blue curve (enrolled users).

The Mann-Whitney U test result for the number of screens visited shows a test statistic of 422652951.0 and a p-value of 0.0. This extremely low p-value indicates a statistically significant difference in the number of screens visited by users who are enrolled in the premium service compared to those who are not enrolled.

The boxplot visualization further illustrates this significant difference:

- 1) Non-Enrolled Users: The distribution is lower with fewer outliers, indicating that non-enrolled users tend to visit fewer screens.
- 2) Enrolled Users: The distribution shows a higher median and more variability, with several outliers indicating that some enrolled users visit many more screens.

Based on the Mann-Whitney U test and the boxplot analysis, we confirm the hypothesis that **users who enroll in the premium service visit a significantly different number of screens compared to those who do not enroll is valid.**

More specifically, users who visit a higher number of screens are more likely to enroll.

Regression Test to verify hypothesis

Features: ['VerifyPhone', 'const', 'remain_screen_list', 'credit_screens_count', 'loan_screens_count', 'Alerts', 'VerifyMobile', 'idscreen', 'VerifyCountry', 'numscreens', 'age', 'VerifyDateOfBirth', 'BankVerification', 'used_premium_feature', 'dayofweek', 'minigame']
Summary:

Summary:

Generalized Linear Model Regression Results					
Dep. Variable:	enrolled	No. Observations:	50000		
Model:	GLM	Df Residuals:	49984		
Model Family:	Binomial	Df Model:	15		
Link Function:	Logit	Scale:	1.0000		
Method:	IRLS	Log-Likelihood:	-24939.		
Date:	Wed, 08 May 2024	Deviance:	49878.		
Time:	20:57:09	Pearson chi2:	5.15e+04		
No. Iterations:	5	Pseudo R-squ. (CS):	0.3221		
Covariance Type:	nonrobust				

	coef	std err	z	P> z	[0.025	0.975]
VerifyPhone	1.2759	0.031	41.528	0.000	1.216	1.336
const	-1.4555	0.048	-30.602	0.000	-1.549	-1.362
remain_screen_list	0.2623	0.006	46.890	0.000	0.251	0.273
credit_screens_count	0.4844	0.012	39.416	0.000	0.460	0.508
loan_screens_count	-0.7242	0.020	-36.009	0.000	-0.764	-0.685
Alerts	-0.9051	0.052	-17.436	0.000	-1.007	-0.803
VerifyMobile	1.2950	0.060	21.674	0.000	1.178	1.412
idscreen	0.3843	0.024	16.120	0.000	0.338	0.431
VerifyCountry	-0.6384	0.034	-18.560	0.000	-0.706	-0.571
numscreens	-0.0161	0.001	-14.999	0.000	-0.018	-0.014
age	-0.0140	0.001	-13.331	0.000	-0.016	-0.012
VerifyDateOfBirth	0.2562	0.025	10.217	0.000	0.207	0.305
BankVerification	0.1754	0.028	6.369	0.000	0.121	0.229
used_premium_feature	-0.1949	0.033	-5.884	0.000	-0.260	-0.130
dayofweek	0.0130	0.005	2.379	0.017	0.002	0.024
minigame	-0.0775	0.037	-2.120	0.034	-0.149	-0.006

AIC: 49909.74572020706

AIC: 49909.74572020706

Fig 8. Regression Analysis Results of the best performing model

We used a forward selection method based on AIC to iteratively select the best features for a binomial Generalized Linear Model model and identify the best model based on the AIC score. The goal was to find the most simplest model that adequately explains the data.


Summary of Regression Analysis Results

1. Model Fit:

Pseudo R-squared (Comparative Fit): 0.3221. suggests a moderate fit of the model.

2. Model Coefficients:

The regression results shown reflect a GLM using a logistic regression approach to predict the likelihood of "enrolled" based on various predictors. The model is robust (non robust covariance type). Key predictors with significant positive effects on enrollment include



"VerifyPhone" and "VerifyMobile," which suggest that verifying user information through these means increases the likelihood of enrollment. On the other hand, "const" and "Alerts" have significant negative effects, indicating that these factors decrease the likelihood of enrollment. The model also includes other predictors like "age" and "BankVerification," which have smaller but still statistically significant impacts.

Key insights of regression analysis:

Hypothesis 1)

The coefficient for 'age' is significant ($p\text{-value} < 0.000$) with a positive coefficient of 0.0101. This suggests that as age increases, the likelihood of enrolling increases slightly. This confirms that there's a relationship between age and enrollment.

Hypothesis 2)

The coefficient for 'numscreens' is -0.0160 with a $p\text{-value} < 0.000$. This suggests that with each additional screen visited, the likelihood of enrollment increases slightly, implying that those who do enroll visit more screens.

Results

In the conducted analysis, we investigated the relationship between user enrollment in a premium service and two key variables: age and the number of screens visited. The findings from the statistical tests provide clear insights:

Hypothesis 1) Age and Premium Enrollment:

The Mann-Whitney U test showed a statistically significant difference in the ages of enrolled versus non-enrolled users, with younger users more likely to enroll in the premium service.

Hypothesis 2) Number of Screens and Premium Enrollment:

The Mann-Whitney U test confirmed a significant difference in the number of screens visited by enrolled versus non-enrolled users, with enrolled users visiting more screens.

These results suggest that younger users and those who interact more with the app by visiting numerous screens are more inclined to subscribe to premium services. The implications of these findings suggest potential areas for targeted marketing and app design improvements aimed at increasing premium subscription rates.

Teammates Results:

Hypothesis 3) Day of the week

This result from the ANOVA test indicates that there is not statistically significant difference in enrollment based on the day of the week, with a p-value of 0.12, which is not significant.

Based on these findings, there is evidence regarding the influence of the day of the week on enrollment decisions. The initial observation from the bar plot suggests no significant effect, and the ANOVA test results also indicate a statistically insignificant difference in enrollment across different days of the week.

Therefore we can interpret that days of the week do not have a statistically significant effect on enrollment status.

Hypothesis 4) Used premium features


In conclusion, the analysis of the association between user behaviors and the likelihood of enrolling in the premium service has provided valuable insights into the factors influencing user decisions. From the investigation into Hypothesis 3, compelling evidence indicates a significant correlation between the usage of premium features during the free trial period and the propensity for enrollment. Despite a marginally lower enrollment rate among active premium feature users, the statistical significance of this association suggests the influential role premium feature engagement plays in shaping user enrollment decisions.

Hypothesis 5) liking features in the app

The finding directly addresses the hypothesis and indicates the following:

1. The point-biserial correlation coefficient between liking a feature and enrollment in the premium version is approximately -0.002, which is very close to zero.
2. The correlation is reported as "non-significant" with a p-value of 0.695.
3. The results suggest that there is a very weak negative correlation between liking a feature and enrollment in the premium version.

Based on this information, we can infer that there is no significant correlation between users liking features in the app and their likelihood of enrolling in the premium service. The



correlation coefficient is very close to zero, and the p-value of 0.695 indicates that the correlation is not statistically significant.

Hypothesis 6) Mini games

The difference in enrollment rates between mini-game engagers and non-engagers was relatively small, statistical tests revealed a significant correlation between mini-game engagement and enrollment likelihood. Regression analysis further quantified this relationship, emphasizing the potential of interactive elements like mini-games in driving user engagement and ultimately influencing enrollment decisions.

Conclusion

The statistical analysis performed on the dataset from a fintech app has given us significant insights into user behavior, preferences and behaviors. Our findings demonstrate that both age and the extent of interaction with the app (as measured by the number of screens visited) significantly correlate with the likelihood of enrolling in premium services. Specifically, younger users and those who are more engaged with the app tend to subscribe to the premium services. The other hypothesis similarly helps us understand the quality of our product, by analyzing users' interaction with the app via looking at the results of other hypotheses tested.

These insights offer actionable recommendations for the fintech company. By focusing on enhancing user engagement through app design and targeting specific demographics more aggressively with marketing campaigns, the company can potentially increase its premium service subscriptions. Additionally looking at interaction of the user with features of the app, such as mini games played, was the user ever enrolled in a subscription, whether the user liked a feature, we can gain insights into what parts of our product needs improvement.

Together, these hypotheses provide a multi-faceted view of the variables that could be leveraged to enhance user acquisition and retention strategies. The insights derived from these analyses suggest actionable strategies for the fintech company. By targeting younger demographics, enhancing interactive elements like mini-games, and promoting the benefits experienced during free trials of premium features, the company can increase its premium service subscriptions. Furthermore, understanding the positive feedback (likes) within the app can help tailor marketing and product development efforts more effectively. This strategic use of data-driven insights is crucial in optimizing the app's design and marketing campaigns, ultimately driving higher conversion rates and user satisfaction in the competitive fintech landscape.

References:

[1]: Data set:

https://www.kaggle.com/datasets/hkhamnakhalid/customers-to-subscription-through-app-behavior/data?select=top_screens.csv

[2]: Mann-Whitney U test https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

[3]: Q-Q plot:

<https://library.virginia.edu/data/articles/understanding-q-q-plots#:~:text=A%20QQ%20plot%20is%20a,truly%20come%20from%20normal%20distributions.>

[4]: Shapiro Wilk test: https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test