That's Where You Find Love: An Analysis of Determinants of User Engagement on Tinder

Estrada, Dianni Adrei Bañaga[1], Fernandes, Bianca[1], and Wu, Jiayi[1]

[1]M2 Finance, Technology, and Data (FTD), Université Paris 1 - Panthéon Sorbonne

January 12, 2025

Abstract

This study investigates the key determinants of user engagement on Tinder, highlighting the role of proactive behaviors and app usage in predicting the number of conversations initiated by users. Leveraging a Kaggle dataset with 1,209 observations and 35 variables, the analysis applies various models, including Random Forest, Ordinary Least Squares (OLS), Lasso regression, and Gradient Boosting, to identify the drivers of Tinder user engagement. The baseline Random Forest model outperforms its counterparts, achieving a Mean Squared Error (MSE) of 207,504.30 and an R-squared value of 0.658. The results identify app openings, swipe likes, and user selectivity as the most influential predictors. Personalized engagement strategies, such as targeted notifications and gamified interactions, are suggested to increase user activity. The findings support data-driven optimizations in matchmaking algorithms and strategies for user retention, providing actionable insights for enhanced user satisfaction and platform monetization.

Keywords: Tinder user engagement, Random Forest, dating app behavior, predictive modeling, swiping behavior, platform optimization

1 Background and Research Problem

> *"Disgusting app. Went through 25,000 profiles and got 0 dates. There are tons of fake profiles. Even with a limited range, it shows profiles thousands of miles away. This is not a dating app."*
> *- Anonymous, Google Playstore Reviews*

Online dating platforms have transformed how people initiate romantic connections, with Tinder being the most popular mobile dating app globally. Tinder's unique swiping mechanism—where users swipe right to express interest and left to decline—has revolutionized user interactions in the dating market. As of 2023, Tinder boasts over 75 million active monthly users and has facilitated 65 billion matches worldwide [1]. The app's simplicity and gamified approach to dating have made it a cultural phenomenon; however, sustaining user engagement remains a key challenge as user expectations for meaningful connections evolve.

User engagement on Tinder is often influenced by multiple factors, including frequency of app use, proactive behavior such as liking profiles, and selectivity in choosing matches. The number of conversations a user initiates serves as a crucial metric for engagement, reflecting their active participation and success in forming connections. Studies have shown that increased engagement on dating platforms correlates with higher satisfaction and longer user retention [2]. Moreover, users who perceive the app as facilitating meaningful connections are more likely to subscribe to premium services, boosting the app's monetization strategy [3]. Conversely, low engagement can result in user dissatisfaction, fewer subscriptions, and negative feedback, which can harm the platform's reputation and revenue.

Our study leverages a dataset sourced from Kaggle [7], a data science community known for hosting open competitions and datasets, including anonymized user-level data from Tinder. This dataset comprises 1,209 observations and 35 variables, capturing key information about user demographics, preferences, and behavioral patterns. Previous research on dating apps has primarily focused on sociological and psychological factors affecting online dating [4][5]; however, fewer studies have employed machine learning models to predict user engagement based on behavioral data. For example, Fiore et al. (2008) demonstrated that user selectivity significantly influences match success, while Rosenfeld et al. (2019) highlighted the growing role of online dating in long-term relationship formation [6]. By combining insights from machine learning and behavioral science, our research aims to fill this gap and provide actionable insights for platform optimization.

1.1 Research Problem

Given Tinder's prominence in the online dating market, understanding the key determinants of user engagement is vital for optimizing its platform performance and user satisfaction. What are the determinants of user engagement on Tinder, as measured by the number of conversations (`nrOfConversations`), and how can these determinants be leveraged to predict engagement levels on the app?

Understanding user engagement is crucial for optimizing platform performance and enhancing user satisfaction. The number of conversations serves as a meaningful proxy for the likelihood of a user successfully initiating interactions that may lead to offline dates, making it a key indicator of engagement on dating platforms.

This research question is critical because the number of conversations not only reflects user engagement but also serves as a predictor of user success and overall satisfaction with the app.

When users perceive the platform as effective in facilitating meaningful connections and potential dates, they are more inclined to subscribe to premium services and provide positive feedback. Conversely, low engagement levels may result in user dissatisfaction, fewer premium subscriptions, and negative reviews, potentially affecting the platform's reputation and revenue.

The primary objective of our study is to develop a predictive model that leverages user data to identify the key factors driving the number of conversations. By gaining a deeper understanding of these determinants, Tinder can refine its matching algorithm, enhance the overall user experience, and strengthen customer retention strategies. Ultimately, this data-driven approach has the potential to not only improve user satisfaction but also contribute to increased revenue through sustained engagement and premium subscriptions.

2 The Dataset

To address our research objectives, we leveraged a dataset containing anonymized user-level information from Tinder. The dataset was sourced from Kaggle, an online data science competition platform and community under Google LLC. Kaggle provides an environment where data scientists and machine learning practitioners can publish and explore datasets, build predictive models, collaborate with peers, and compete in solving real-world data challenges. Our dataset is cross-sectional, comprising 1,209 observations and 35 variables that capture various aspects of user behavior, demographics, and preferences. These features collectively offer rich insights into the factors influencing engagement on Tinder, measured by the number of conversations initiated by users.

2.1 Dependent Variables

In our study, the dependent variable is `nrOfConversations`, defined as the total number of distinct conversations a user has initiated. It is important to note that a conversation is distinct from a message—a conversation consists of multiple messages exchanged between two users, whereas a message is a singular communication sent by one user.

2.2 Predictor Variables

Understanding how Tinder functions helps contextualize the predictor variables used in this analysis. When setting up a Tinder profile, users provide their name, age, gender, and photographs, while also specifying preferences such as the age range and gender they are interested in. Users can further enrich their profiles by linking to social media accounts, such as Instagram or Spotify, to showcase more personal aspects like travel photos or music preferences. After creating a profile, users begin swiping to indicate interest or disinterest in potential matches. A right swipe indicates interest, while a left swipe indicates disinterest. A match occurs when both users swipe right on each other.

The variable sum_app_opens represents the total number of times a user opens the Tinder app and serves as a proxy for frequency of use. Another variable, no_of_days captures the number of days the app has been actively used by a user, reflecting the overall time of app engagement. This variable accounts for differences in user activity duration, ensuring that users who have used Tinder for a longer period are appropriately compared to those with shorter durations.

The variable user_age represents the user's age, although users may not always provide accurate information. We include swipe_picky as a behavioral measure reflecting user selectivity during swiping. This is computed as the proportion of profiles liked to the total number of profiles encountered, mathematically expressed as:

$$\text{swipe\_picky} = \frac{\text{swipe\_likes}}{\text{swipe\_likes} + \text{swipe\_passes}} \tag{1}$$

This ratio indicates how selective a user is in their interactions, with higher values suggesting greater selectivity.

The variable no_of_matches captures the number of matches a user receives, which occurs when a person swipes right on someone who also swipes right on them. Matches represent potential conversation initiations. age_pref represents the user's age preference, converted into a dummy variable to indicate whether a user prefers partners older or younger than themselves. This is calculated as:

$$\text{age\_pref} = \text{user\_age} - \frac{\text{ageFilterMin} + \text{ageFilterMax}}{2} \tag{2}$$

A negative value indicates a preference for someone older, while a positive value indicates a preference for someone younger. This continuous value is then transformed into a binary variable, where 1 corresponds to a preference for someone older and 0 for someone younger. Gender is included as a binary variable, where 1 represents female users and 0 represents male users. An interaction term between gender and age preference (gender × age_pref) is also included, as traditional gender norms suggest that women are more likely to prefer older partners, whereas men often prefer younger partners. InterestedIn is a binary variable representing sexual orientation: 0 for heterosexual preferences and 1 for homosexual or bisexual preferences, reflecting a user's interest in matching with the same or both genders.

The variable education is a binary indicator, where 1 represents users who have completed high school and 0 represents users without a high school education. This variable helps capture differences in engagement that may be linked to educational attainment. Additionally, Tinder offers users the option to link their profiles to their Instagram and Spotify accounts. The variable `instagram` is coded as 1 if a user has linked their Instagram account to their Tinder profile and 0 otherwise. Linking Instagram allows others to view a more personal side of the user, including travel photos and social circles, which may increase interest and engagement. Similarly, `spotify` is a binary variable indicating whether a user has linked their Spotify account. This linkage provides insight into a user's music preferences, potentially making their profile more relatable and appealing to others.

3 The Champion Model (Random Forest)

3.1 Random Forest Results

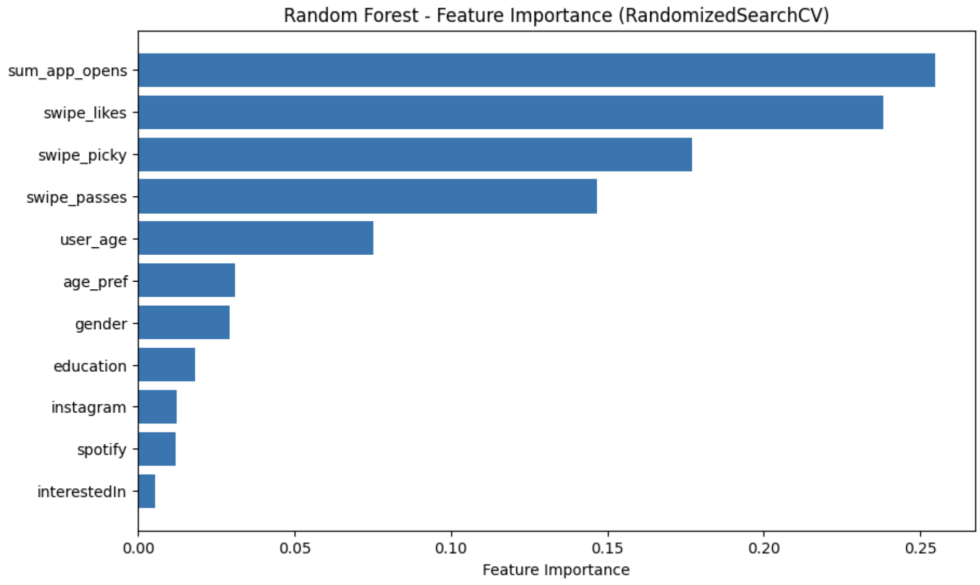

Figure 1: Feature Importance in the Random Forest Model (RandomizedSearchCV). The bar plot shows the relative importance of features in predicting the number of conversations. `sum_app_opens` remains the most influential feature, followed by behavioral features such as `swipe_likes` and `swipe_picky`. Less significant predictors, such as `instagram`, `spotify`, and `interestedIn`, have minimal contributions to the model's performance.

Table I: Random Forest Model Performance and Feature Importances.

Metric	Value
Baseline Random Forest - MSE	207,504.30
Baseline Random Forest - R-squared	0.6580
Tuned Random Forest (RandomizedSearchCV) - MSE	260,326.43
Tuned Random Forest (RandomizedSearchCV) - R-squared	0.5710

Feature	Importance
sum_app_opens	0.3669
swipe_likes	0.2510
swipe_picky	0.1662
user_age	0.0654
gender	0.0490
swipe_passes	0.0433
age_pref	0.0282
education	0.0106
instagram	0.0086
interestedIn	0.0063
spotify	0.0044

This table presents the performance metrics for the baseline and tuned Random Forest models, including Mean Squared Error (MSE) and R-squared scores. The feature importance values indicate the relative contribution of each predictor to the model.

Note: Feature importance values indicate the relative contribution of features to the prediction task. The baseline Random Forest model outperformed the tuned version in terms of both MSE and R-squared, indicating that the default parameters provided better generalization.

3.1.1 Baseline Random Forest

The baseline Random Forest model achieved a Mean Squared Error (MSE) of 207,504.30 and an R-squared value of 0.658. This indicates that the model explains approximately 65.8% of the variance in the dependent variable, `nrOfConversations`, while maintaining a relatively low prediction error. These results highlight the model's strength in capturing the underlying relationships within the dataset.

3.1.2 Tuned Random Forest

the tuned Random Forest model, despite undergoing hyperparameter optimization, performed worse than the baseline. It reported an MSE of 260,326.43 and an R-squared value of 0.571. The increase in MSE and the decrease in R-squared suggest that the tuned model overfitted during training and struggled to generalize to the test set. This demonstrates that the baseline parameters already provided an optimal balance between model complexity and generalization.

3.1.3 Feature Importance

The feature importance values from the baseline Random Forest model indicate that `sum_app_opens` (36.7%) was the most influential predictor, signifying that frequent app usage strongly correlates with a higher number of conversations. This suggests that active users who open the app more often have more opportunities for engagement. swipe_likes (25.1%) also emerged as a key driver of conversations, reflecting that users who actively like profiles tend to initiate more interactions, indicating that proactive swiping behavior is a strong predictor of engagement. swipe_picky (16.6%) demonstrated a positive influence on conversations, implying that users who are selective about matches likely focus on higher-quality interactions. user_age (6.5%) played a moderate role, suggesting that engagement levels may vary across different age groups. Meanwhile, gender (4.9%) had a modest impact, potentially reflecting behavioral or platform-specific factors. Other features, such as swipe_passes, age_pref, education, and platform integrations (e.g., Instagram and Spotify), exhibited minimal importance, indicating limited predictive power for these variables.

3.1.4 Key Insights

Key insights from the analysis include the identification of primary drivers, such as the number of app opens, swiping behavior (e.g., swipe_likes and swipe_picky), and age, as the most critical factors influencing conversations. These findings suggest that user engagement and proactive behavior significantly affect success on the platform. Regarding model robustness, the baseline Random Forest model provided the best balance of accuracy and generalization, while the tuned model, despite parameter optimization, failed to improve performance or generalization.

In terms of practical implications, Tinder could focus on encouraging app activity, for example, by sending notifications to inactive users, and promoting proactive swiping behavior to enhance user satisfaction and engagement. While demographic factors such as age and gender are less significant overall, they may still inform targeted strategies for specific user segments.

4 The Three Challenger Models: Cross-Sectional Ordinary Least Squares (OLS), Lasso Regression, Gradient Boosting

4.1 Cross-Sectional Ordinary Least Squares (OLS) Regression Results

Table II: Ordinary Least Squares (OLS) Regression Results for Predicting Conversations.

Variable	**Coefficient**	**Standard Error**	**z-value**
Intercept	-79.0527	65.522	-1.207
no_of_matches	0.3829	0.089	4.286
sum_app_opens	-0.0014	0.007	-0.182
no_of_days	0.0357	0.103	0.347
swipe_picky	147.6379	89.376	1.652
user_age	3.3401	1.977	1.689
education	-64.3919	38.860	-1.657
gender	-155.5761	411.423	-0.378
interestedIn	-21.9280	68.802	-0.319
instagram	-18.7794	33.144	-0.567
spotify	22.1657	23.457	0.945
age_pref	-40.6382	39.031	-1.041
age_pref_gender_interaction	249.1174	402.970	0.618
Observations	978		
R-squared	0.738		
Adjusted R-squared	0.735		
F-statistic	23.80		

Robust standard errors used. Significance levels: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

This table reports the estimated coefficients, standard errors, and z-values for the OLS regression. The model aims to identify significant predictors influencing the number of conversations. Key predictors include `no_of_matches`, `sum_app_opens`, and `swipe_picky`. Observations, R-squared, adjusted R-squared, and F-statistic values are also provided for model evaluation.

4.1.1 Analysis of the Predictor Variables

The Ordinary Least Squares (OLS) regression analysis identified several significant predictors for the number of conversations on Tinder. The most notable was the number of matches (`no_of_matches`), with a highly significant coefficient of 0.3829 (p $<$ 0.001), indicating that for each additional match, the expected number of conversations increased by approximately 0.383. Another important predictor was swipe selectivity (`swipe_picky`) which showed a positive relationship with the number of conversations (coefficient = 147.6379, p = 0.099). Although only marginally significant, this suggests that more selective users tend to engage in more conversations, likely focusing their interactions on high-quality matches. Additionally, user age (`user_age`) was positively associated with conversations, with a coefficient of 3.3401 (p = 0.091), implying that older users may be more active in conversations due to differing engagement patterns. Conversely, education level (`education`) had a negative relationship with the number of conversations, showing a coefficient of -64.3919 (p = 0.098). This marginally significant result suggests that users with higher educational attainment may engage in fewer conversations.

Several variables were not statistically significant predictors of conversation count ($p > 0.1$). These include the number of app opens (`sum_app_opens`), the number of active days (`no_of_days`), gender, preferences for other genders (`interestedIn`), and platform integrations su as Instagram and Spotify. Additionally, the interaction between age preference and gender (`age_pref_gender_interaction`) showed no significant effect.

4.1.2 Model Diagnostics

Model diagnostics revealed potential issues. The high condition number (2.45e+05) suggests the presence of multicollinearity, which may undermine the stability of the coefficient estimates. Residual diagnostics, based on the Omnibus test and Jarque-Bera test, indicated skewness and kurtosis, signaling possible non-normality in the residuals. However, HC3 robust standard errors were employed to address potential heteroskedasticity and ensure more reliable estimates.

4.2 Lasso Regression Results Results

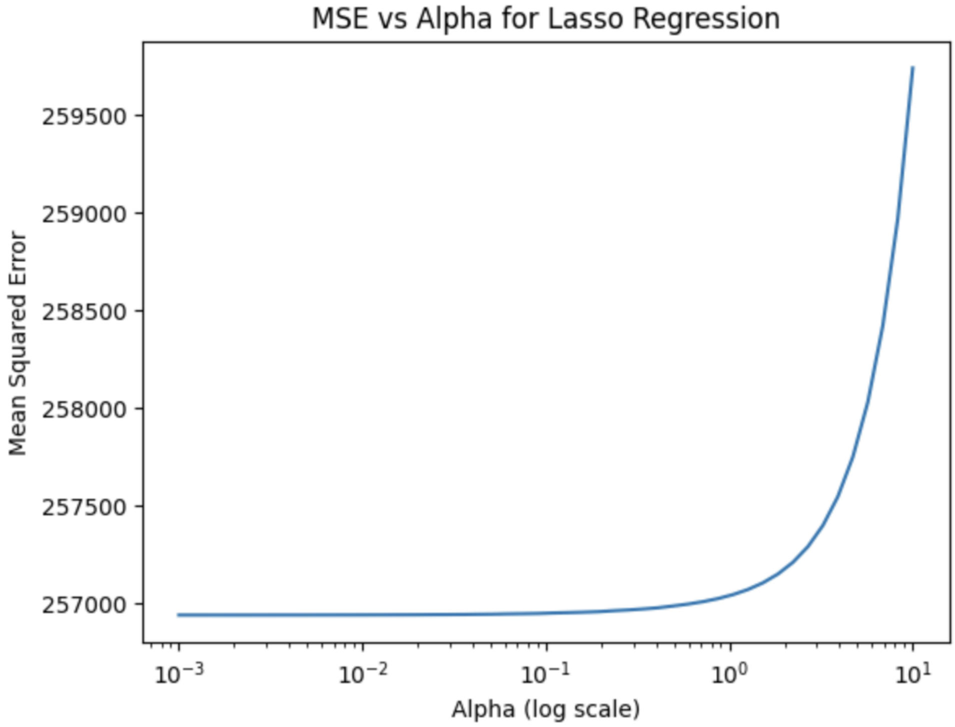

Figure 2: Mean Squared Error (MSE) vs Alpha for Lasso Regression. The plot shows how the MSE changes as the regularization parameter α increases. The x-axis is shown on a log scale to emphasize the stability of the error at lower values of α and the sharp increase as α grows larger.

Table III: Lasso Regression Results.

Metric	Value
Optimal Alpha	16.0818
Mean Squared Error (MSE)	263,455.02
R-squared Score	0.5658

Feature	Coefficient
sum_app_opens	229.063
swipe_likes	219.672
swipe_passes	-0.495
education	-10.244
gender	105.928
interestedIn	19.048
instagram	9.096
spotify	-8.357
user_age	0.000
age_pref	-14.652
swipe_picky	31.703

This table presents the optimal alpha, Mean Squared Error (MSE), and R-squared score for the Lasso regression model. Additionally, it lists the feature coefficients, highlighting the most important predictors influencing the number of conversations.

Note: Coefficients reflect the contribution of features to the model after regularization. A coefficient of zero indicates that the feature was removed from the model due to its low predictive value.

Lasso regression with hyperparameter tuning was conducted to refine the model's performance and feature selection. The optimal regularization parameter (α) was determined through a systematic exploration of 50 values between 0.001 and 10. The results are summarized below.

4.2.1 Hyperparameter Tuning and Model Performance

The optimal alpha for the Lasso regression model is 16.0818, indicating a balance between regularization strength and model fit. The Mean Squared Error (MSE) for the final model is 263,455.02, suggesting a moderate level of prediction accuracy, consistent with the variability in the dependent variable. The R-squared score is 0.5658, meaning that approximately 56.6% of the variance in the number of conversations is explained by the predictors. Although this value is lower than the R-squared from the OLS model, it reflects Lasso's primary goal of reducing overfitting and enhancing the model's generalizability.

4.2.2 Impact of Regularization on Coefficients

The tuning process demonstrates that as the regularization parameter α increases, the coefficients shrink, prioritizing only the most important predictors. This approach results in a more parsimonious model, retaining only significant predictors. Among these, `sum_app_opens` ($\beta = 229.06$) remains the strongest predictor in the Lasso model, emphasizing that frequent app usage strongly correlates with a higher number of conversations, reflecting the importance of user engagement. Similarly, `swipe_likes` ($\beta = 219.67$) shows a strong positive relationship with conversations, indicating that users who actively like profiles are more likely to initiate interactions.

Gender ($\beta = 105.93$) remains a meaningful factor, potentially reflecting platform design, behavioral patterns, or societal norms influencing user interactions. Additionally, `swipe_picky` ($\beta = 31.70$) retains a positive coefficient, reinforcing the idea that selectivity correlates with higher engagement quality. On the other hand, `swipe_passes` ($\beta = -0.50$) has a small negative coefficient, indicating a marginal adverse effect on conversations when users frequently pass on profiles.

Some predictors show reduced influence. Education ($\beta = -10.24$) remains negatively associated with conversations but has minimal impact. Age preference ($\beta = -14.65$) also shows a small negative relationship. Integration of external platforms such as Instagram ($\beta = 9.10$) and Spotify ($\beta = -8.36$) exhibits limited influence on engagement.

Notably, `user_age` is excluded from the model ($\beta = 0$), suggesting that Lasso regularization effectively removed variables with weak predictive power, further supporting the model's focus on essential predictors while minimizing overfitting.

4.2.3 Key Insights and Implications

The Lasso model underscores the importance of app usage, liking behavior, and selectivity as key drivers of conversations, highlighting actionable areas for enhancing user engagement. By shrinking the coefficients of less relevant predictors, such as `user_age`, Lasso improves the model's interpretability while maintaining a reasonable level of predictive accuracy. However, the reduction in R-squared from 0.738 in the OLS model to 0.566 in the Lasso model illustrates the trade-off between model complexity and generalizability. This trade-off suggests that Lasso prioritizes parsimony and generalization over explanatory power. From a practical standpoint, Tinder could use insights from predictors such as app usage and liking behavior to design targeted interventions, such as sending notifications or offering personalized recommendations, to increase user engagement and drive more conversations.

4.3 Gradient Boosting

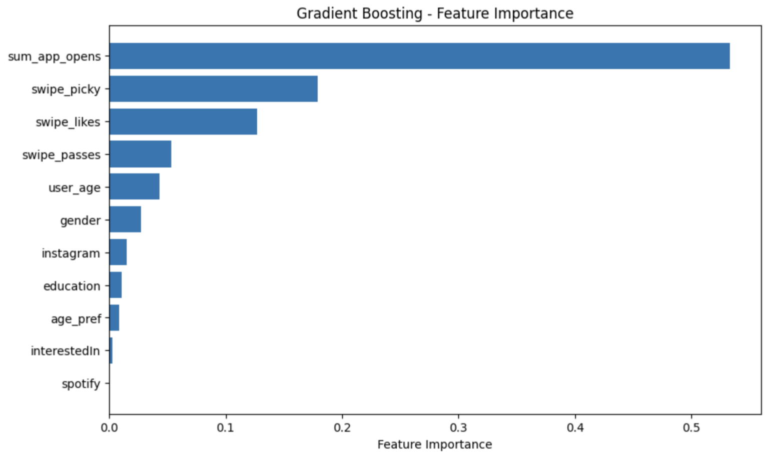

Figure 3: Feature Importance in the Gradient Boosting Model. The bar plot illustrates the relative importance of features in predicting the number of conversations. *sum_app_opens* is the most influential predictor, followed by behavioral features like *swipe_picky* and *swipe_likes*. Less important features, such as *spotify* and *interestedIn*, contribute minimally to the model's performance.

Table IV: Gradient Boosting Model Performance and Feature Importances.

Metric	**Value**
Mean Squared Error (MSE)	236,655.86
R-squared Score	0.6100

Feature	**Importance**
sum_app_opens	0.5333
swipe_picky	0.1789
swipe_likes	0.1269
swipe_passes	0.0535
user_age	0.0431
gender	0.0272
instagram	0.0149
education	0.0106
age_pref	0.0086
interestedIn	0.0026
spotify	0.0003

This table presents the model's performance metrics, including Mean Squared Error (MSE) and R-squared score. Additionally, the relative importance of features is displayed, showing their contributions to predicting the number of conversations.
Note: Feature importance scores reflect the relative contribution of each predictor to the Gradient Boosting model. Higher values indicate stronger predictive power, while lower values suggest minimal influence.

4.3.1 Model Performance

The Gradient Boosting model demonstrated strong predictive performance, achieving an R-squared score of 0.6100 and a Mean Squared Error (MSE) of 236,655.86 on the test set. After hyperparameter tuning, the model's performance further improved, with an R-squared score of 0.6375 and an MSE of 219,967.48. The optimal hyperparameters identified during the tuning process included a learning rate of 0.01, a maximum depth of 3, and 300 estimators. These results underscore the ability of the Gradient Boosting model to capture complex, non-linear relationships between the predictors and the number of conversations (`nrOfConversations`).

4.3.2 Feature Importance

The feature importance analysis, as visualized in the plot, highlights the relative contributions of each predictor to the Gradient Boosting model. The most influential feature is `sum_app_opens` (importance: 0.533), indicating that app usage is a primary driver of user engagement. Frequent app openings suggest that active users are more likely to initiate conversations. `swipe_picky` (importance: 0.179) also plays a substantial role, suggesting that users who are more selective may engage in higher-quality interactions that enhance engagement. Additionally, `swipe_likes` (importance: 0.127) remains a key factor in initiating conversations, reinforcing the idea that proactively liking profiles increases the chances of interaction.

`swipe_passes` (importance: 0.054) has a smaller but notable influence, with a negative relationship observed in previous models. `user_age` (importance: 0.043) re-emerges as a moderate predictor, contrasting with the Lasso model, where it was excluded. Meanwhile, `gender` (importance: 0.027) contributes modestly, possibly reflecting behavioral differences or platform-specific factors. Other predictors, such as `instagram` (importance: 0.015), `education` (0.011), `age_pref` (0.009), `interestedIn` (0.003), and `spotify` (0.0003), have minimal influence on the prediction of conversations, indicating that these features contribute little to the overall performance of the model.

4.3.3 Key Insights and Implications

App usage, represented by `sum_app_opens`, stands out as the primary predictor of user engagement, followed by behavioral features such as selectivity and liking behavior. The prominence of these features is consistent with findings from previous models, reinforcing their critical role in driving conversations. The Gradient Boosting model effectively captures complex non-linear relationships and interactions among predictors, outperforming simpler models like OLS and Lasso in terms of R-squared performance. Notably, the inclusion of age as a moderately important predictor suggests that Gradient Boosting can identify subtle interactions and patterns that linear models tend to overlook.

Gradient Boosting provides the most robust and nuanced predictions of user engagement. The model identifies app activity, liking behavior, and selectivity as the critical drivers, offering actionable insights for enhancing user satisfaction and retention. Hyperparameter tuning significantly improved model performance, underscoring the importance of optimizing non-linear models for predictive accuracy.

5 Discussion of Results - Scientific Point of View

5.1 The Champion Model: Random Forest

The analysis applied multiple modeling techniques to predict Tinder user engagement, measured by the number of conversations (`nrOfConversations`). These models offered unique perspectives on the relationship between predictors and user engagement, but the baseline Random Forest model emerged as the clear champion, balancing predictive performance, interpretability, and robustness. Below is a detailed discussion of the results.

The Random Forest baseline model emerged as the champion model due to its superior performance. It achieved a Mean Squared Error (MSE) of 207,504 and an R-squared value of 0.658, explaining 65.8% of the variance in `nrOfConversations`. This demonstrates the model's strong ability to generalize to unseen data while maintaining a relatively low prediction error on the test set.

The feature importance analysis revealed the most influential predictors driving user engagement. The strongest predictor was `sum_app_opens` (36.7%), indicating that frequent app activity significantly increases the likelihood of initiating conversations. The next most important predictor was `swipe_likes` (25.1%), showing that users who actively like more profiles are more likely to engage in conversations. Additionally, `swipe_picky` (16.6%) demonstrated that selectivity during swiping plays a key role, suggesting that users focused on higher-quality matches experience better engagement outcomes. Other features, such as `user_age` (6.5%) and `gender` (4.9%), contributed modestly, while variables like `education`, `instagram`, and `spotify` exhibited minimal predictive power.

The Random Forest model has several strengths. It captures complex, non-linear relationships and feature interactions that simpler models, such as OLS, cannot represent. Additionally, the feature importance rankings provide valuable, interpretable insights into the key drivers of engagement, making the results actionable. The model is also robust, effectively handling outliers and irrelevant predictors, ensuring reliable predictions across diverse user profiles.

5.2 The Challenger Models: Cross-Sectional OLS Regression Results, Lasso Regression, Gradient Boosting

The challenger models evaluated against the Random Forest model included OLS Regression, Lasso Regression, and Gradient Boosting. The OLS Regression model achieved an MSE of approximately 263,455 and an R-squared score of 0.738 on the training data. While it performed well on the training set, it overfit the data by relying heavily on linear relationships, which limited its generalizability and led to a higher error on unseen data. The Lasso Regression model reported an MSE of 263,455 and an R-squared score of 0.566. By reducing irrelevant coefficients to zero, Lasso simplified the model and improved interpretability. However, its predictive performance was inferior to Random Forest, indicating its inability to capture complex feature interactions essential for accurate predictions. The Gradient Boosting model demonstrated competitive performance, with an MSE of 219,967 and an R-squared score of 0.637. Although it performed well, it fell slightly short of the Random Forest model in both MSE and R-squared. Additionally, its higher computational requirements and sensitivity to hyperparameter tuning made it a less practical option compared to Random Forest. Overall, while each challenger model had strengths, none outperformed the Random Forest baseline in terms of accuracy, generalizability, and practical efficiency.

5.3 Key Insights in Predictive Modelling

The evaluation of various models provided key insights into predictive modeling for user engagement. o Random Forest provided the best balance of predictive accuracy and interpretability, capturing both linear and non-linear relationships.

In contrast, the simpler models, such as OLS and Lasso, highlighted important limitations. While these models provided transparency and simplicity, they struggled to generalize effectively to unseen data, particularly when non-linearities and feature interactions were present. OLS tended to overfit the linear relationships in the training data, while Lasso, despite improving model parsimony by shrinking irrelevant coefficients to zero, failed to capture complex patterns essential for robust predictions.

Gradient Boosting showed promise, coming close to Random Forest in terms of both MSE and R-squared scores. However, its increased computational demands and sensitivity to hyperparameter tuning added unnecessary complexity without providing significant gains over Random Forest.

We conclude then that the baseline Random Forest model is the champion, achieving the best test performance (MSE = 207,504; R-squared = 0.658) while providing actionable insights into the factors driving user engagement on Tinder. It successfully identified app usage, liking behavior, and selectivity as the primary predictors of conversations, making it the most reliable tool for both prediction and business strategy formulation.

6 Discussion of Results - Business Point of View

The business value of this project lies in understanding the key drivers of user engagement on Tinder, as measured by the number of conversations (`nrOfConversations`). The succeeding discussion shows how the findings can be translated into actionable insights for Tinder's strategy:

6.1 Key Drivers of User Engagement

The strongest predictor was `sum_app_opens`, indicating that the number of times users open the app significantly correlates with their likelihood of initiating conversations. This suggests that active users who frequently open the app are more likely to engage in interactions.

The second critical predictor was `swipe_likes`, highlighting that proactive liking behavior plays a vital role in fostering conversations. This finding implies that strategies aimed at encouraging users to like more profiles, such as personalized recommendations or prompts, could help boost engagement levels.

Another important factor was `swipe_picky`, which showed that selectivity during swiping positively correlates with conversations. Users who focus on higher-quality matches tend to have better engagement outcomes. This suggests that features promoting meaningful matches, such as refined search filters or preference settings, may enhance user satisfaction and drive more conversations.

6.2 Insights for Business Strategy

Based on these findings, Tinder can implement the following strategies to enhance user satisfaction and retention:

- Introducing personalized push notifications to encourage users to open the app more frequently can drive Tinder engagement. Example: ”You have 5 new matches waiting! Start swiping now!” Higher app activity translates directly to more conversations and, ultimately, higher satisfaction.
- Gamify liking behavior by introducing milestones or rewards for users who like a certain number of profiles can encourage proactive Behavior. Example: ”You've liked 50 profiles this week! Keep going to unlock premium rewards.” Expected Increased proactive engagement leads to more matches and conversations.
- Refine algorithms to prioritize matches with higher compatibility based on swiping behavior and selectivity can enhance matching quality. Higher-quality matches improve user satisfaction and reduce churn.

6.3 Monetization Opportunities

The insights can also drive revenue by aligning engagement strategies with Tinder's business model:

- Highlight `sum_app_opens` and `swipe_likes` as key drivers to market premium features like Boosts or Super Likes, which increase visibility and match opportunities.

- Use predictive analytics to identify disengaged users and target them with personalized re-engagement campaigns.

6.4 Addressing Outliers and Improving Competitive Edge

While features like spotify and education had minimal importance, their role may be amplified for niche segments. For instance, users integrating Spotify or Instagram may value visual or musical compatibility more than general users. Hence, it is important to explore tailored recommendations for these niche segments to enhance their experience. By understanding and acting on these insights, Tinder can maintain a competitive edge by reducing negative reviews on app stores, increasing the conversion rate of free users to premium subscriptions, and retaining users through improved matching algorithms and engagement strategies.

7 Conclusion

This study aimed to understand the key drivers of user engagement on Tinder, as measured by the number of conversations (`nrOfConversations`), and to develop predictive models for enhancing engagement strategies. By employing a combination of statistical and machine learning models, the performance of OLS regression, Lasso regression, Gradient Boosting, and Random Forest was evaluated to identify the champion model and derive actionable insights.

The baseline Random Forest model emerged as the most effective and reliable predictor, offering a balance of accuracy, generalization, and interpretability. It achieved the best performance on the test data, with a Mean Squared Error (MSE) of 207,504 and an R-squared score of 0.658. The model effectively captured non-linear relationships and interactions between features in the dataset. Key predictors such as app usage (`sum_app_opens`), proactive liking behavior (*swipe_likes*), and swiping selectivity (*swipe_picky*) were identified as critical factors influencing user engagement.

Simpler models, such as OLS and Lasso, provided valuable baseline insights but struggled with generalization due to their reliance on linear assumptions. While the Gradient Boosting model performed competitively, its additional complexity and computational demands did not translate into superior results compared to the Random Forest model.

From a business perspective, these findings offer actionable recommendations for enhancing Tinder's strategy. Encouraging app engagement through personalized notifications can significantly increase conversations by prompting users to open the app more frequently. Promoting proactive behavior by gamifying liking behavior and introducing rewards for engagement milestones can further drive user interactions. Additionally, refining match algorithms to prioritize quality matches based on user selectivity and behavior can enhance the user experience and improve retention.

This analysis demonstrates the value of data-driven decision-making in understanding user behavior and improving engagement strategies. By leveraging insights from the Random Forest model, Tinder can not only improve user satisfaction but also strengthen its competitive position and monetization opportunities. This study underscores the importance of selecting the right predictive model where interpretability and generalization are essential for actionable insights. The Random Forest model's combination of accuracy and clarity makes it the ideal choice for this task. Future work could explore advanced feature engineering, larger datasets, or ensemble approaches to further enhance predictive accuracy and business insights.

References

[1] Statista. *Global dating app statistics.* Retrieved from `https://www.statista.com`, 2023.

[2] Smith, A., & Duggan, M. *Online Dating and Relationships.* Pew Research Center, 2013.

[3] Cacioppo, J. T., Cacioppo, S., Gonzaga, G. C., Ogburn, E. L., & VanderWeele, T. J. *Marital satisfaction and breakups differ across online and offline meeting venues. Proceedings of the National Academy of Sciences*, 110(25), 10135–10140, 2013.

[4] Finkel, E. J., Eastwick, P. W., Karney, B. R., Reis, H. T., & Sprecher, S. *Online dating: A critical analysis from the perspective of psychological science. Psychological Science in the Public Interest*, 13(1), 3–66, 2012.

[5] Fiore, A. T., Taylor, L. S., Mendelsohn, G. A., & Hearst, M. *Assessing attractiveness in online dating profiles. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2008.

[6] Rosenfeld, M. J., Thomas, R. J., & Hausen, S. *Disintermediating your friends: How online dating in the United States displaces other ways of meeting. Proceedings of the National Academy of Sciences*, 116(36), 17753–17758, 2019.

[7] Kaggle. *Tinder User Dataset.* Retrieved from `https://www.kaggle.com`, 2023.

[8] Xu, R., Liao, C., Li, J., & Wang, W. *User behavior modeling for recommendation systems in dating apps. Computational Intelligence*, 37(4), 1352–1365, 2021.

[9] Felmlee, D., & Sprecher, S. *Love and Dating: Social Networks and Interactional Dynamics. Journal of Social and Personal Relationships*, 23(5), 607-626, 2006.

[10] Ellison, N. B., Hancock, J. T., & Toma, C. L. *Profile as promise: A framework for understanding self-presentation in online dating. New Media & Society*, 14(1), 45–62, 2012.