Tinderlytics Report

"Unveiling the Patterns and Dynamics of Online Dating on Tinder"

Sergio Ambelis Diaz
Becca Nika
Alexander Moran
Herman Contreras

Introduction

Tinder is one of the leading platforms in the rapidly changing world of digital romance, revolutionizing the way people communicate and connect. An extensive examination of Tinder's user habits, demographics, and social dynamics is provided by this initiative, Tynderlitics. Our goal was to unravel user interaction complexities and understand how factors like age, gender, occupation, and education level influence dating preferences and patterns on Tinder.

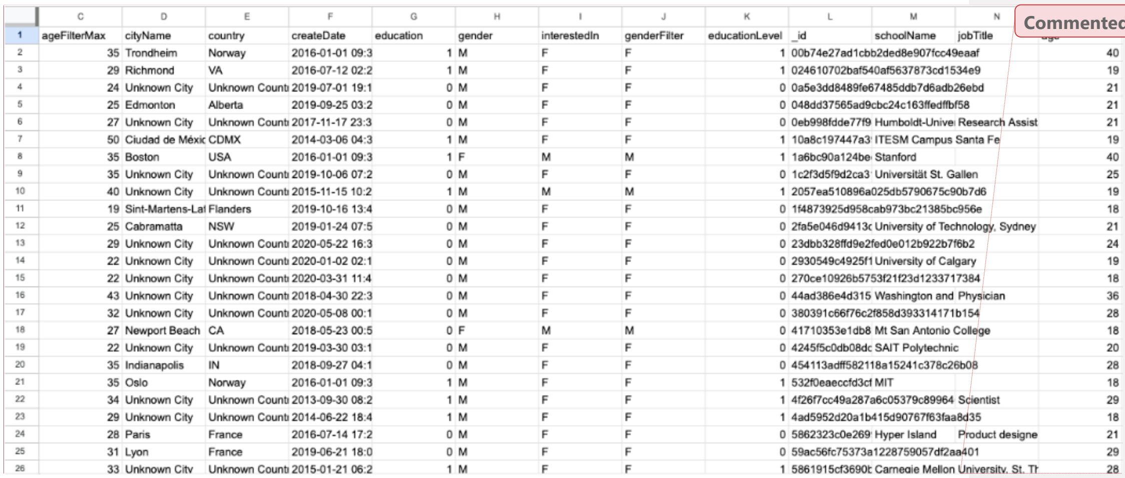
Commented [1]: By: Sergio Ambelis Diaz

Project Overview and Data Description

We began with two different datasets: a .json file of 536 MB and a .csv file of 360 MB, rich in quantitative and categorical data. These include metrics such as app opens, message exchanges, likes, passes, and matches, along with user-specific details like age, gender, occupation, and educational background. Our analysis is rooted in a comprehensive data preparation process, including the transformation of json objects into data frames, cleaning of user data, and critical handling of outliers and missing values.

Commented [2]: By: Sergio Ambelis Diaz

Commented [3]: By: Herman Contreras



	C	D	E	F	G	H	I	J	K	L	M	N	
1	ageFilterMax	cityName	country	createDate	education	gender	interestedIn	genderFilter	educationLevel	_id	schoolName	jobTitle	age
2	35	Trondheim	Norway	2016-01-01 09:3	1	M	F	F	1	00b74e27ad1cbb2ded8e907fcc49eaaf			40
3	29	Richmond	VA	2016-07-12 02:2	1	M	F	F	1	024610702baf540af5637873cd1534e9			19
4	24	Unknown City	Unknown Counti	2019-07-01 19:1	0	M	F	F	0	0a5e3dd8489fe67485ddb7d6adb26ebd			21
5	25	Edmonton	Alberta	2019-09-25 03:2	0	M	F	F	0	048dd37565ad9cbc24c163ffedffbf58			21
6	27	Unknown City	Unknown Counti	2017-11-17 23:3	0	M	F	F	0	0eb998fdde77f9	Humboldt-Unive	Research Assist	21
7	50	Ciudad de Méxic	CDMX	2014-03-06 04:3	1	M	F	F	1	10a8c197447a3	ITESM Campus Santa Fe		19
8	35	Boston	USA	2016-01-01 09:3	1	F	M	M	1	1a6bc90a124be	Stanford		40
9	35	Unknown City	Unknown Counti	2019-10-06 07:2	0	M	F	F	0	1c2f3d5f9d2ca3	Universität St. Gallen		25
10	40	Unknown City	Unknown Counti	2015-11-15 10:2	1	M	M	M	1	2057ea510896a025db5790675c90b7d6			19
11	19	Sint-Martens-Lat	Flanders	2019-10-16 13:4	0	M	F	F	0	1f4873925d958cab973bc21385bc956e			18
12	25	Cabramatta	NSW	2019-01-24 07:5	0	M	F	F	0	2fa5e046d9413c	University of Technology, Sydney		21
13	29	Unknown City	Unknown Counti	2020-05-22 16:3	0	M	F	F	0	23dbb328ffd9e2fed0e012b922b7f6b2			24
14	22	Unknown City	Unknown Counti	2020-01-02 02:1	0	M	F	F	0	2930549c4925f1	University of Calgary		19
15	22	Unknown City	Unknown Counti	2020-03-31 11:4	0	M	F	F	0	270ce10926b5753f21f23d1233717384			18
16	43	Unknown City	Unknown Counti	2018-04-30 22:3	0	M	F	F	0	44ad386e4d315	Washington and	Physician	36
17	32	Unknown City	Unknown Counti	2020-05-08 00:1	0	M	F	F	0	380391c66f76c2f858d393314171b154			28
18	27	Newport Beach	CA	2018-05-23 00:5	0	F	M	M	0	41710353e1db8	Mt San Antonio College		18
19	22	Unknown City	Unknown Counti	2019-03-30 03:1	0	M	F	F	0	4245f5c0db08dc	SAIT Polytechnic		20
20	35	Indianapolis	IN	2018-09-27 04:1	0	M	F	F	0	454113adff582118a15241c378c26b08			28
21	35	Oslo	Norway	2016-01-01 09:3	1	M	F	F	1	532f0eaeccfd3cf	MIT		18
22	34	Unknown City	Unknown Counti	2013-09-30 08:2	1	M	F	F	1	4f26f7cc49a287a6c05379c89964		Scientist	29
23	29	Unknown City	Unknown Counti	2014-06-22 18:4	1	M	F	F	1	4ad5952d20a1b415d90767f63faa8d35			18
24	28	Paris	France	2016-07-14 17:2	0	M	F	F	0	5862323c0e269	Hyper Island	Product designe	21
25	31	Lyon	France	2019-06-21 18:0	0	M	F	F	0	59ac56fc75373a1228759057df2aa401			29
26	33	Unknown City	Unknown Counti	2015-01-21 06:2	1	M	F	F	1	5861915cf3690b	Carnegie Mellon University, St. Th		28

We merged these datasets, linking user information with their matching data on _id. This involved a strategic use of 'outer' and 'left' joins to ensure integration without losing crucial data points. Amidst this merging, we tackled NaN values and outliers with precision, transforming dates from strings to datetime objects and filling missing categorical data with placeholder values.

Commented [4]: By: Sergio Ambelis Diaz

Among this data exploration, we address important hypothetical questions that shape the online dating landscape:

- **Gender Distribution and Preferences**: How does gender distribution among users influence their preferences between males and females?
- **User Activity Frequency**: What patterns emerge in the frequency of user activity on Tinder?
- **Professional and Educational Dominance**: Which professions and education backgrounds are most prevalent among Tinder's user base?
- **Age Preference Dynamics**: What are the minimum and maximum age preferences per user, and how are they influenced by gender?
- **Age Group Analysis**: What is the median age group among Tinder users?

Commented [5]: By: Sergio Ambelis Diaz

Our project's scope, while initially broader, has been fine-tuned to focus on these key areas, excluding aspects like conversation dynamics and frequency of user activity to sharpen our analytical lens. We aim to provide users with valuable insights into Tinder's platform, assessing whether it truly delivers on its promises or falls short in the face of data-driven scrutiny.

The data, sourced from third-party provider swipestats.io, comprises over 1,000 Tinder user profiles. However, we encountered challenges such as missing information, skewed data, and the presence of fake profiles. Despite these hurdles, our analysis endeavors to paint a comprehensive picture of the Tinder universe.

Commented [6]: By: Sergio Ambelis Diaz

Cleaning the data involved removing NaN and missing data from our results. Being careful not to introduce bias results and allowing for transparency. Additionally, we analyzed data types in columns such as user birth dates and account creation to convert to datetime. We excluded irrelevant columns, such as Instagram and Spotify details. We made to only remove data that wouldn't impact our findings. Furthermore, we performed calculations, including determining the users age. The cleaning was performed after data was imported and merged (csv and json) into one file. The unique key to our data was heavily reliant on _id which was given to each user.

Commented [7]: By Alexander Moran

Commented [8]: By: Sergio Ambelis Diaz

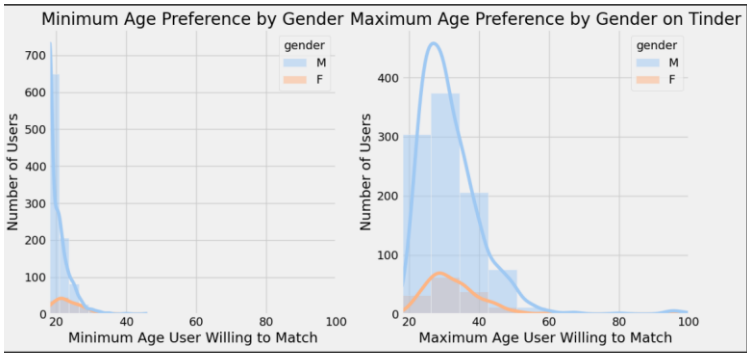
```
clean_user_data = user_data.copy()
clean_matches_data = matches_data_filtered.copy()
tinder = pd.merge(clean_user_data, clean_matches_data, on='_id',
how='left')
print(tinder.head(2))
```

```
clean_user_data = tinder.copy()
```

Data Processing and Exploratory Data Analysis (EDA)

In looking at how people on Tinder choose who they might want to date, we noticed something interesting about age. Everyone agreed that they want to match with people who are at least 18 years old, but not much younger. As for the older age limit, people were more flexible. They were okay with matching with people up to around 40 years old. This was true for both men and women. So, it looks like people are pretty strict about not wanting to date anyone too young, but they're more open to dating older people.

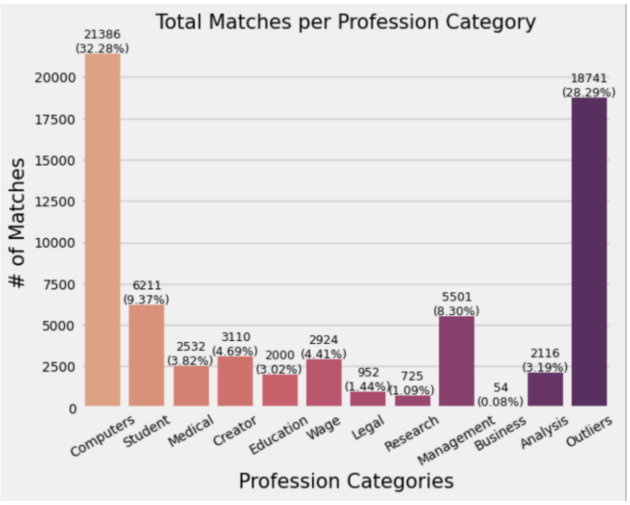
Commented [9]: By: Sergio Ambelis Diaz



Another factor in age is that the majority of both male and female users are young adults, with median ages of 23 for males and 22 for females respectively. This finding suggests that Tinder's user base skews towards a younger audience, a detail that not only informs us about the typical age range of users but also potentially influences user preferences and matching patterns on the platform. Understanding this age distribution is crucial for contextualizing behaviors within the app, as the priorities and dating preferences of younger adults may distinctly shape the interactions and connections that occur within Tinder's digital environment.

Commented [10]: By: Sergio Ambelis Diaz

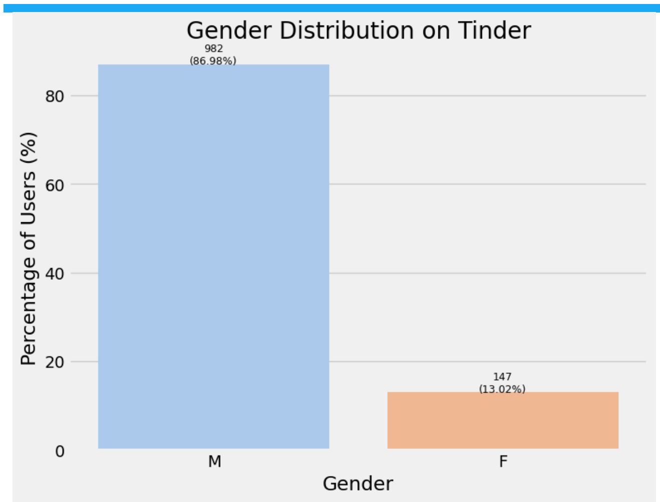
Another interesting factor that we found when studying our data was that there was a much higher quantity of users within the tech field, making up 56% of our users. Every other type of profession that we categorized was much less, with only jobs we considered to be outliers either because they weren't real or were in a different language getting even remotely close at nearly 30%. Going one step further from that, we calculated the amount of matches that each user received based on their profession and created the following visualization.



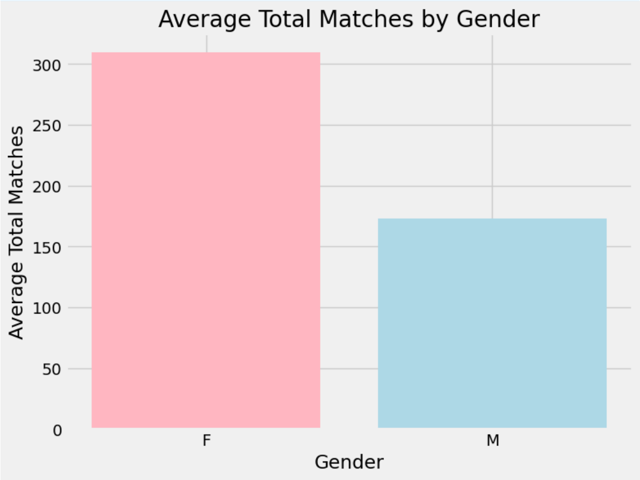
Even though the users with a job related to computers made up more than half of the total users, they only make up 32% of the total matches. Still a much higher percentage than the others, minus the outliers which was much closer at 28%, but not as much as one might have assumed. Analyzing the profession distribution is also important because it will not only allow us to understand the variety of skill sets and professions that users of tinders possess, but we can also utilize the match data to see whether a users job affects their success on Tinder. A user's background can also change the course of their conversations with other users, and given how important professions are in this society, we can also analyze how they may affect the conversations people have had and whether or not they were eventually ghosted.

Commented [11]: By: Becca Nika

Exploring another fascinating aspect of our data is examining the gender distribution done by Alexander Moran. Our dataset reveals a notable imbalance, with a higher percentage of male users (87%) compared to female users (13%), giving us a male to female ratio of 7:1, indicating a substantial skew towards males on the Tinder platform.



Adding onto the insights of the gender distribution, Alexander Moran investigated the average total matches for both male and female users made. Our findings show that on average, female users receive twice as many matches as male users.



This insight in particular is intriguing given the skew male presence in our dataset (7:1 ratio male to female). Despite the heavy male user representation, female users receive 2x the average number of matches. In the end, we see that a gender imbalance and a match discrepancy exist.

Commented [12]: by Alexander Moran

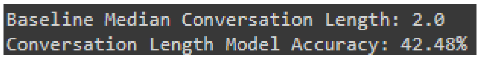
Machine Learning Analysis

In our Machine Learning Analysis section, Sergio Ambelis Diaz employed the K-Nearest Neighbors (KNN) algorithm. The objective was to analyze user engagement on Tinder by predicting the median length of conversations. We chose features pertaining to user interaction intensity: the percentage of one-message conversations, which reflects initial engagement, and the longest conversation length, which suggests sustained interest.

The baseline median conversation length, computed as 2.0, highlighted a general tendency towards brief conversation exchanges. With a model accuracy of 42.48%, our findings suggest that while one-message conversations are common, factors contributing to longer conversations might be more complex. This insight reveals that while initial engagement on Tinder is high, sustaining long conversations may require more than just user demographics and initial interaction patterns, indicating a need for incorporating more fine grained behavioral data into our analysis to fully understand and enhance user engagement strategies on the platform.

Commented [13]: By: Sergio Ambelis Diaz

Commented [14]: By: Sergio Ambelis Diaz

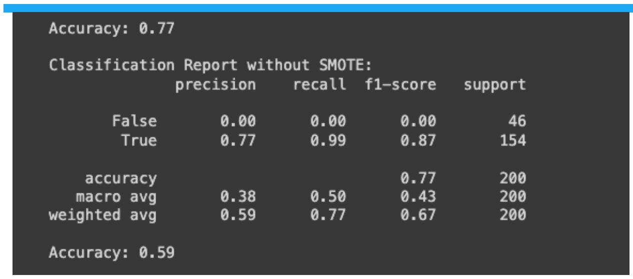


```
Baseline Median Conversation Length: 2.0
Conversation Length Model Accuracy: 42.48%
```

Another ML analysis that was used by Alexander Moran was a Linear Regression model to gain insight on the connection between the age of Tinder users and their total matches. The main focus was to predict if a user has matches or no matches. Two models were built, one without SMOTE and the other with SMOTE.

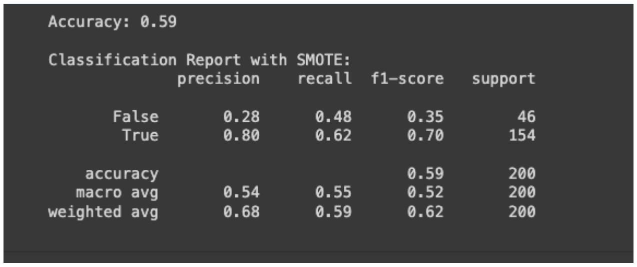
The data included data on total matches and age. To prepare the data, any rows with missing values were handled by dropping them to make sure the set of data was complete. The target variable was “isMatch” based on whether “totalMatches” was greater than 0. For the feature variable, the users ‘age’ was used. The feature was then scaled using the StandardScaler.

In the first model without SMOTE, the accuracy was 0.77, showing that it correctly predicted whether there was a match about 77% of the time. However, the classification report has trouble identifying users with no matches (False class), which showed low precision and recall.



```
Accuracy: 0.77

Classification Report without SMOTE:
              precision    recall  f1-score   support

       False       0.00      0.00      0.00        46
        True       0.77      0.99      0.87       154

    accuracy                           0.77       200
   macro avg       0.38      0.50      0.43       200
weighted avg       0.59      0.77      0.67       200

Accuracy: 0.59
```

In order to address the imbalance, SMOTE was used in the second model. The data was resampled to oversample the True class. When that was applied, the accuracy decreased to 0.59, suggesting that while SMOTE helped balance the classes, the overall performance of the model was not high.



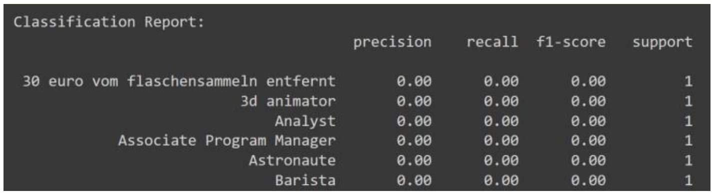
```
Accuracy: 0.59

Classification Report with SMOTE:
              precision    recall  f1-score   support

       False       0.28      0.48      0.35        46
        True       0.80      0.62      0.70       154

    accuracy                           0.59       200
   macro avg       0.54      0.55      0.52       200
weighted avg       0.68      0.59      0.62       200
```

To improve this analysis, we would need to include additional features other than age.

Commented [15]: by Alex

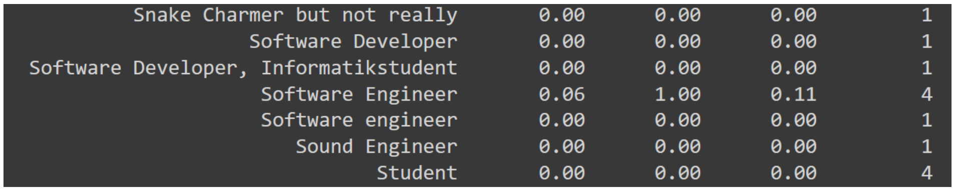
A third analysis by Becca Nika that was attempted also utilized a Linear Regression model to find a connection between the total matches users received and their profession. There was only one model built. The data was prepared similarly as the previous analysis with the values that were null. The feature variable utilized the totalMatches variable and was scaled using the StandardScalar as well.

However, this model was not as successful as the previous model. The accuracy of this model ended up being 6%, with the precision, recall, and f1-score ending up being 0, with the exception of one case: when the profession was ‘Software Engineer’.



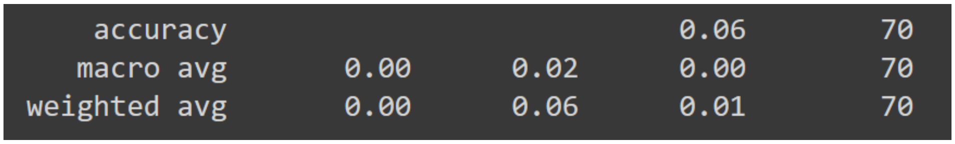
```
Classification Report:
                                 precision    recall  f1-score   support

 30 euro vom flaschensammeln entfernt       0.00      0.00      0.00         1
                          3d animator       0.00      0.00      0.00         1
                              Analyst       0.00      0.00      0.00         1
            Associate Program Manager       0.00      0.00      0.00         1
                           Astronaute       0.00      0.00      0.00         1
                              Barista       0.00      0.00      0.00         1
```

…



```
         Snake Charmer but not really      0.00      0.00      0.00         1
                   Software Developer      0.00      0.00      0.00         1
 Software Developer, Informatikstudent      0.00      0.00      0.00         1
                    Software Engineer      0.06      1.00      0.11         4
                    Software engineer      0.00      0.00      0.00         1
                       Sound Engineer      0.00      0.00      0.00         1
                              Student      0.00      0.00      0.00         4
```

…



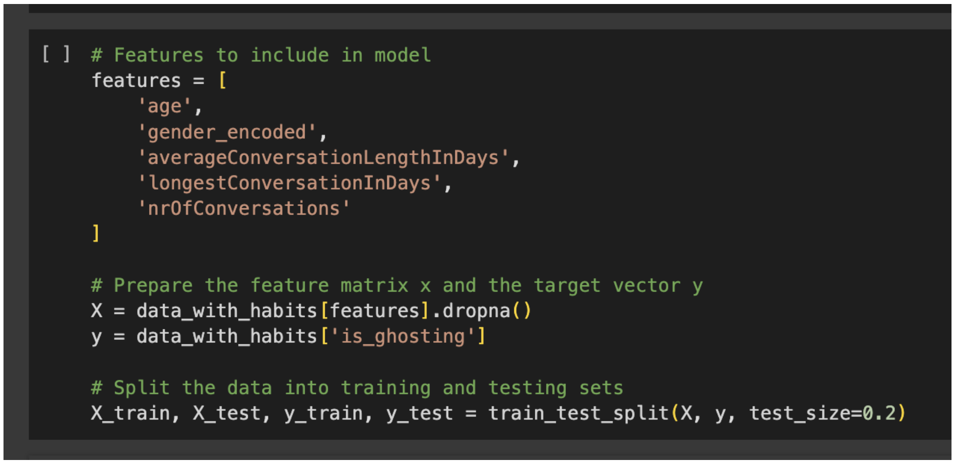
```
    accuracy                           0.06        70
   macro avg       0.00      0.02      0.00        70
weighted avg       0.00      0.06      0.01        70
```

The reason for this could be because many of these professions were unique to each other and thus ended up producing incredibly small numbers when run through the model. One solution for attempting to solve this could be to combine the professions into categories like had been done earlier in the EDA section for visualizations, but that also has the possibility of skewing results.

Commented [16]: By: Becca Nika

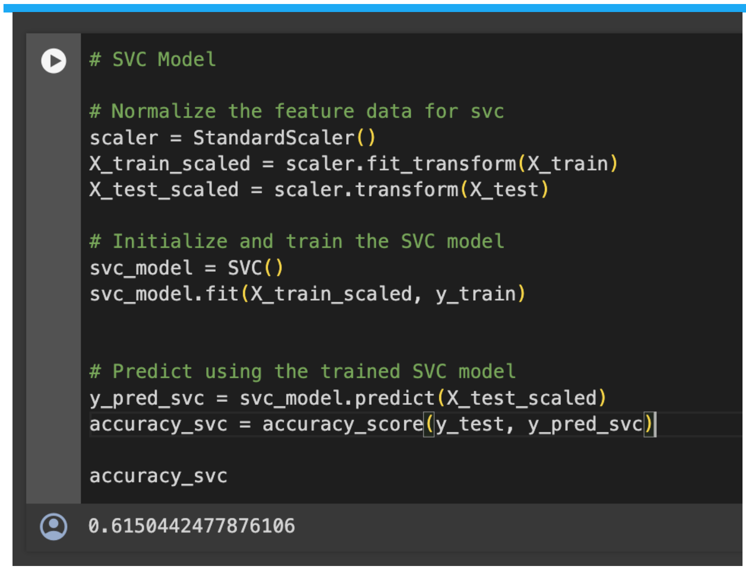
Another machine learning analysis was a predictive model that uses different physical attributes to predict if a hypothetical user is likely to ghost.

Commented [17]: By Herman Contreras



```
[ ]  # Features to include in model
     features = [
         'age',
         'gender_encoded',
         'averageConversationLengthInDays',
         'longestConversationInDays',
         'nrOfConversations'
     ]

     # Prepare the feature matrix x and the target vector y
     X = data_with_habits[features].dropna()
     y = data_with_habits['is_ghosting']

     # Split the data into training and testing sets
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

This list of attributes can be easily altered but we chose to go with common behavioral habits that were consistently filled out and had few missing entries.



```
# SVC Model

# Normalize the feature data for svc
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Initialize and train the SVC model
svc_model = SVC()
svc_model.fit(X_train_scaled, y_train)


# Predict using the trained SVC model
y_pred_svc = svc_model.predict(X_test_scaled)
accuracy_svc = accuracy_score(y_test, y_pred_svc)

accuracy_svc
```

```
0.6150442477876106
```

For this algorithm we used an SVC model and we found that we can predict whether a user is likely to ghost with 61% accuracy.

Reflection and Insights

Several challenges stood out prominently. The first was the data itself; dealing with a dataset that contains both quantitative and categorical information required a lot of cleaning and preparation to ensure accuracy. Missing or skewed data, such as location details, school data and the presence of fake profiles, posed hurdles that had to be overcome through rigorous data validation processes. Such a process included replacing NaN with unknown data to not cause errors, also we did not want to exclude missing data to represent bias in our result, therefore we chose to be strict in our findings.

Commented [18]: By: Sergio Ambelis Diaz

We discovered that user behavior on Tinder is not solely dictated by the straightforward metrics of age or number of conversations but is influenced by a complex interplay of user engagement levels, gender preferences, and even conversation quality. Our machine learning analyses, specifically the K-Nearest Neighbors algorithm, offered a glimpse into the predictability of conversation lengths, which is a significant factor in determining user engagement.

Commented [19]: By: Sergio Ambelis Diaz

When utilizing the Linear Regression model, we were able to determine two separate models (one using SMOTE and the other not) to try and predict how age could affect the total matches a user gained. We tried a similar model in predicting how a user's profession could affect their total matches, but that didn't go as well as the models using age. There is still some improvement needed in attempting to create a good model to accurately predict how age and professions could affect the matches a user receives.

Commented [20]: By: Becca Nika

Conclusions and Next Steps

In conclusion our team, Tinderlytics, uncovered interesting insights into Tinder user behavior, revealing the dynamics of online dating. Notably, age preferences emerged as an important factor, with users leaning towards matching within a specific age range. It's interesting to see that while there's a consensus on wanting matches who are at least 18 years old, users appear more flexible about their upper maximum age preference. Users are willing to match with others who are up to 40 years old, this trend is observed through both male and female users. The median age of Tinder users are 23 for males and 22 for females, further showcasing the influence the role of age has in shaping user interactions on Tinder.

Occupation also proved to be a key finding in the Tinder experience, particularly users in the tech sector. The tech field accounted for about 32%, which has one of the higher match rates right next to the outlier occupations. This suggests that the type of occupation does influence the way users connect on Tinder.

Additionally, our analysis highlighted significant gender imbalance on the platform, with female users receiving double the average matches despite a skewed male representation. This finding underscores the complexity of gender dynamics in the online landscope. In summary, fostering inclusivity, ensuring fair representation, and understanding the preferences of users is important to enhancing the statistician of the online dating experience.

Commented [21]: by alex

Links:

- Link to Report
- Link to Github Repo: amoran9/tinderlytics (github.com)
- Filtered CSV
- Orginal Data:
  - CSV
  - JSON
- Python Script:
- Entire Drive: CS418_FinalProject - Google Drive

References:

- "Match Group Can Get Away with Acquiring 25 Dating Sites and Counting." *Yahoo Finance*, Yahoo, https://finance.yahoo.com/news/match-group-can-get-away-acquiring-25-dating-sites-counting-151306438.html. Accessed 10/06/2023.
- "Match Group." *Yahoo Finance*, Yahoo, https://tinyurl.com/mry7fwbt. Accessed

10/14/2023.

- "SwipeStats: Tinder Data." SwipeStats, https://swipestats.io. Accessed 10/12/2023.
- "Tinderlytics Team Contributions." *Trello*, https://trello.com/b/VLTVTv0l/tinderlytics. Accessed 11/01/2023.
- "Matplotlib: Visualization with Python." Matplotlib, https://matplotlib.org/stable/api/index.html.
- "Seaborn Tutorial." Seaborn, https://seaborn.pydata.org/tutorial.html.