Abstract

The data for this project represents how our users use Tinder. We collect data on every person a user swipes on to the number of messages that are exchanged. The purpose of this study to analyze data, and answer some questions such as the how many male or female use system, on average how many messages are sent and how many received etc. I will answer following questions in this study. This study shows that on average basis, females sent less swipes, likes and messages as compared to males and vice versa in receiving. Users of age group 18 to 28 years are main target audience.

Research Questions (RQs)

1) Basic Exploratory Data Analysis
2) Who sent/ receive more swipes on average by following
    a. Which country users?
    b. Male or Female?
    c. Android or IOS users?
    d. Day wise average swipes sent by users?
3) Who sent/ receive more average likes on average by following
    a. Which country users?
    b. Male or Female?
    c. Android or IOS users?
    d. Day wise average swipes sent by users?
4) Who sent/ receive more average number of messages on average by following
    a. Which country users?
    b. Male or Female?
    c. Android or IOS users?
    d. Day wise average swipes sent by users?
5) What is probability of having match for users knowing the features such as gender, swipes sent/ received, likes sent/ receive, messages sent/ receive etc., using machine Learning Algorithm. (**Classification Problem**)
6) Next time knowing above features, how many matches are possible in login using Machine Learning Algorithm (**Regression Problem**)
7) Recommendation System: Recommending users which are most likely to pair-up/ match with each other based on available features such as likes, swipes and messages send/ received. (**using machine learning k-NN algorithm**)

RQ1: Data Description

There are 35780 rows and 14 columns/ potential features in given dataset. The dataset contains no missing values. Following is the data dictionary of each of the column.

- **user_id** : unique identifier for each user
- **day** : the day in which data was collected on the user
- **age** : the user's age on that day
- **country** : the country where a user resided on that day
- **gender** : the user's gender (female or male)

- **device_type** : the type of device the user was using on that day (android or ios)
- **is_active_user** : whether the user opened the app that day (1 = Yes, 0 = No)
- **swipes_sent** : the number of people the user swiped on that day
- **swipes_received** : the number of people that swiped on the user on that day
- **likes_sent** : the number of people the user swiped right (liked) on that day
- **likes_received** : the number of people who swiped right on (liked) a user on that day
- **matches** : the number of people the user matched with on that day. matches occur when two people swipe right on (liked) each other
- **messages_sent** : the number of messages the user sent on that day
- **messages_received** : the number of messages the user received on that day

This data is of 1 year and 1 month.

Data Types

Table 1: Data Types

Feature	Data Type	Pre-processing Done
user_id	Integer	-
day	Object	Convert to date time datatype Extract Year, Month, Day
age	Integer	There was age issue for same users
country	Object	-
gender	Object	-
device_type	Object	-
is_active_user	Integer	-
swipes_sent	Integer	-
swipes_received	Integer	-
likes_sent	Integer	-
likes_received	Integer	-
matches	Integer	-
messages_sent	Integer	-
messages_received	Integer	-

Data Unique Values for each Column

Feature	Unique Values
user_id	2198
day	31
age	47
country	5
gender	2
device_type	2
is_active_user	2

swipes_sent	1030
swipes_received	1216
likes_sent	242
likes_received	637
matches	84
messages_sent	276
messages_received	193

Issue(s) With Data

1) Invalid Values

There is some issue with some users, like they have same user ID, country and gender but different age, which is not possible. So i will remove such duplicates to make our analysis clear. There are 150 such users which contain duplicate/ erroneous ages for users. Solution is to ignore one of the values.

	user_id	age	country	gender	device_type
197	11	45	Argentina	Male	android
198	11	46	Argentina	Male	android
931	53	26	Indonesia	Female	ios
932	53	27	Indonesia	Female	ios
1185	77	25	Spain	Male	android
1186	77	26	Spain	Male	android
1211	78	24	Indonesia	Male	ios
1223	78	24	Indonesia	Male	android
1335	84	37	Indonesia	Male	android
1336	84	38	Indonesia	Male	android

2) Outliers/ Anomalies

There were some values which might be outliers in some features as shown.

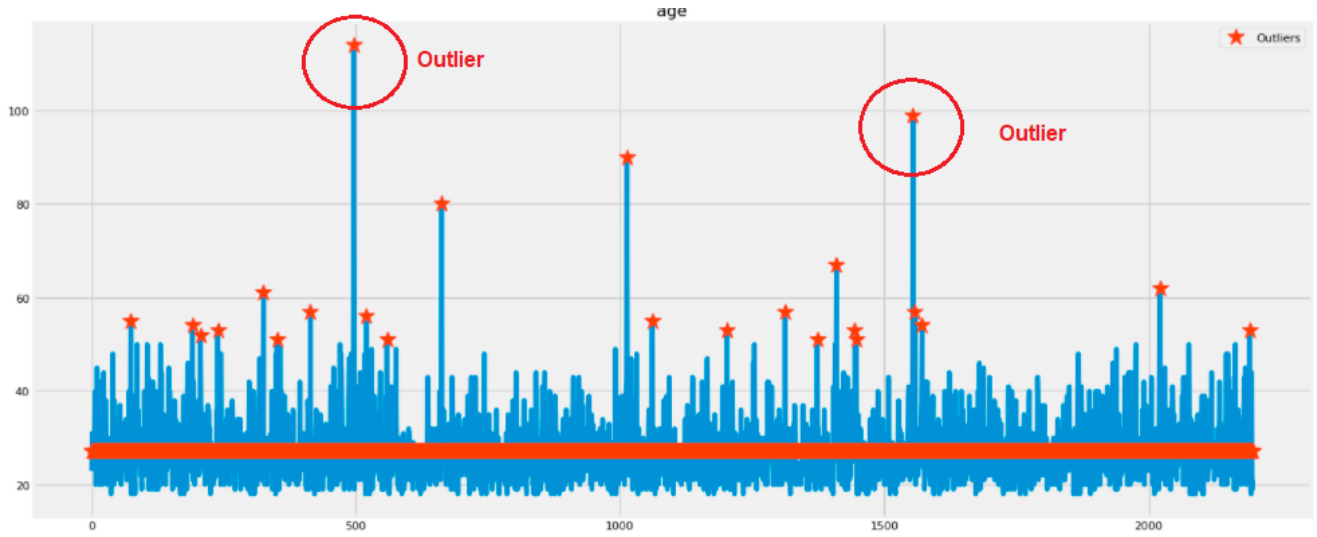

Fig 1: Values such as 114 and 89 in age are outliers

Table 3: Country Wise Users Distribution

Country	Percentage of Users
Argentina	18.79%
Indonesia	16.15%
Italy	14.38%
Spain	28.25%
Thailand	22.43%

Country	Male	Female
Argentina	15.02%	24.22%
Indonesia	21.26%	8.78%
Italy	12.94%	16.44%
Spain	25.50%	32.22%
Thailand	25.27%	18.33%

Table 4: Gender Wise Users Distribution

Country	Percentage of Users
Female	40.95%
Male	59.05%

Table 5: Device Type Wise Users Distribution

Country	Percentage of Users
Android	58.37%
IOS	41.63%

Table 6: Day Wise Users Distribution

Country	Percentage of Users
Friday	16.12%
Monday	12.97%
Saturday	12.79%
Sunday	12.91%
Thursday	16.15%
Tuesday	13.07%
Wednesday	16.00%

Table 7: Status (Active/ Not Active) Users Distribution

Country	Percentage of Users
0 (Not Active)	39.95%
1 (Active)	60.05%

Most of the users are the age bracket 18 to 32 years.

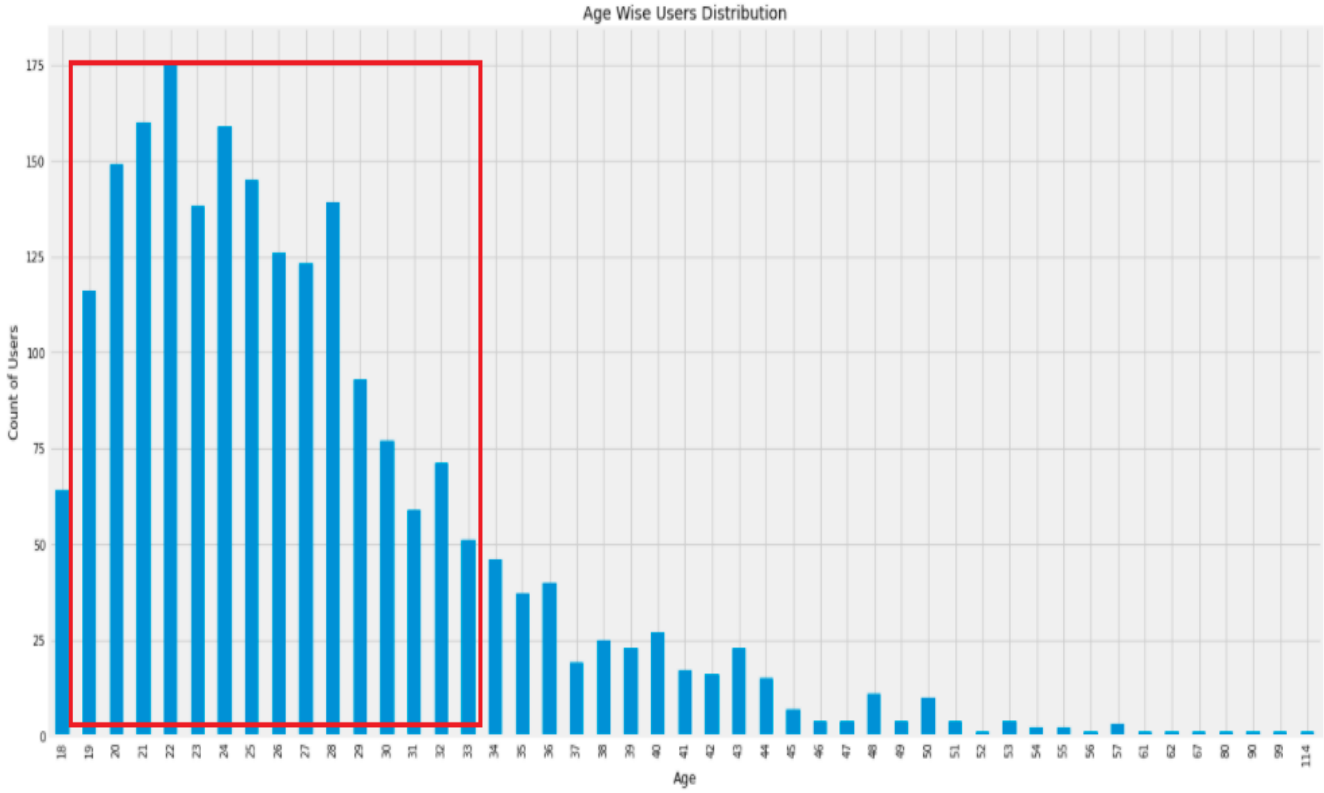

Fig 2: Age Wise User Distribution

RQ2: Swipes Sent/ Received

- Male (65 on average) sent swipes more frequently than Female (62 on average)

- Alternatively, Male (64 on average) received swipes more frequently than Female (125 on average)

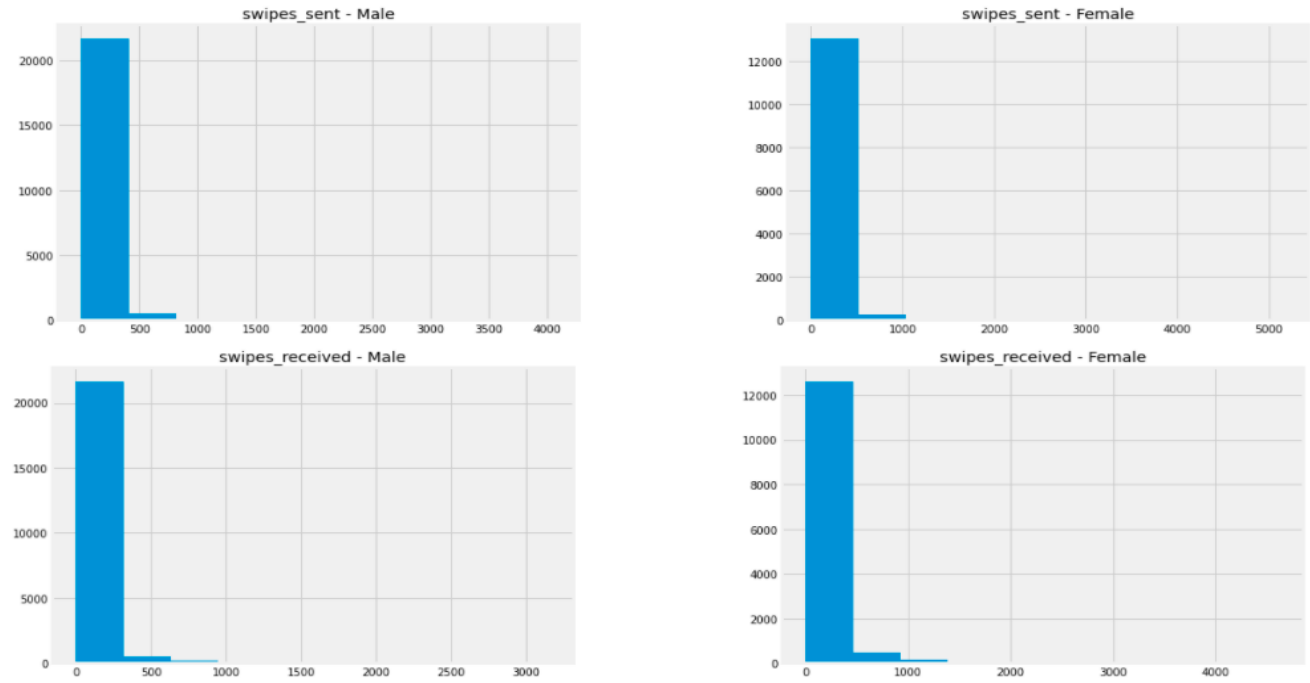

Fig 3: Gender Wise Swipes Sent/ Received

RQ3: Likes Sent/ Received

- Male (12 on average) sent likes more frequently than Female (6 on average).
- Male (10 on average) received likes more frequently than Female (46 on average)

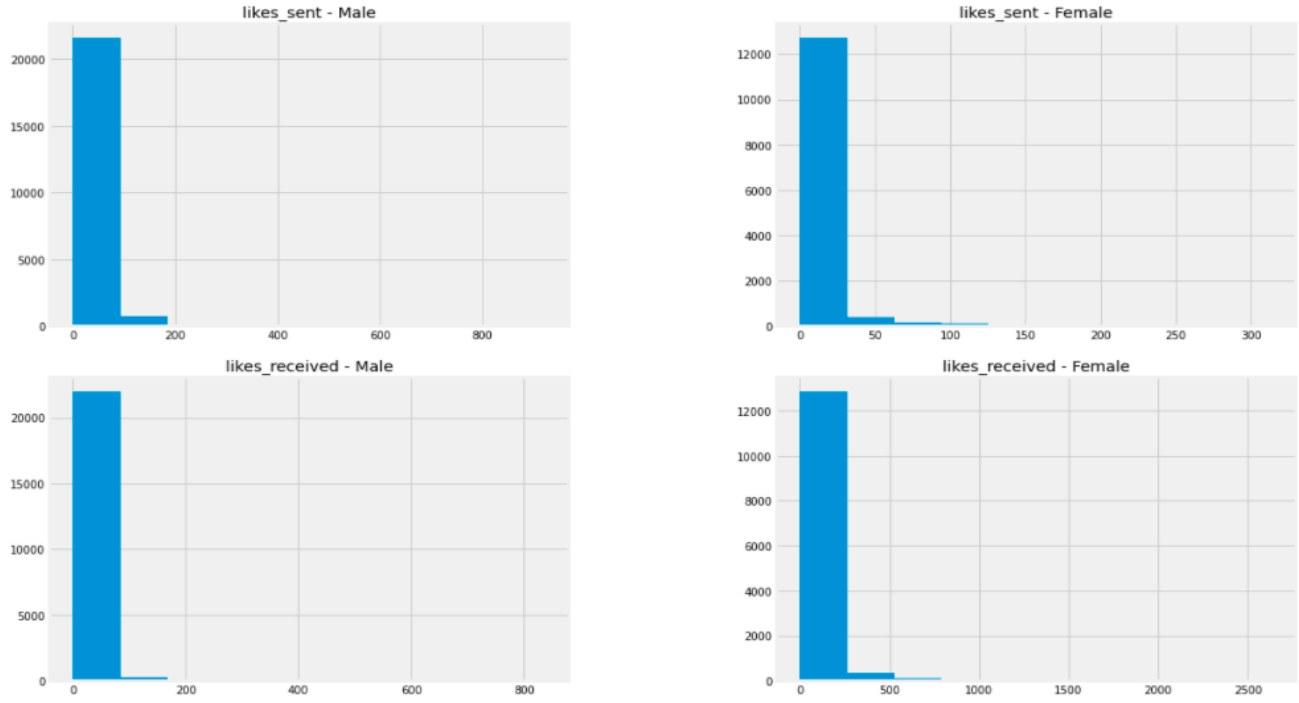

Fig 4: Gender Wise Likes Sent/ Received

RQ4: Messages Sent/ Received

- Male (7 on average) sent messages more frequently than Female (6 on average).
- Male (7 on average) received messages same as Female (7 on average)

	country	swipes_sent	swipes_received	likes_sent	likes_received	matches	messages_sent	messages_received
0	Argentina	58.0	84.0	8.0	27.0	1.0	5.0	5.0
1	Indonesia	56.0	78.0	10.0	18.0	2.0	7.0	8.0
2	Italy	71.0	119.0	11.0	45.0	2.0	6.0	7.0
3	Spain	61.0	90.0	9.0	25.0	1.0	7.0	7.0
4	Thailand	75.0	75.0	11.0	13.0	1.0	6.0	6.0

Fig 5: Country Wise Average Swipes/ Likes/ Messages - Sent/ Received

	device_type	swipes_sent	swipes_received	likes_sent	likes_received	matches	messages_sent	messages_received
0	android	51.0	84.0	9.0	23.0	1.0	6.0	7.0
1	ios	81.0	91.0	11.0	26.0	2.0	6.0	7.0

Fig 6: Device Type Wise Average Swipes/ Likes/ Messages - Sent/ Received

	Day_Name	swipes_sent	swipes_received	likes_sent	likes_received	matches	messages_sent	messages_received
0	Friday	61.0	84.0	10.0	24.0	1.0	6.0	6.0
1	Monday	65.0	86.0	10.0	23.0	1.0	6.0	7.0
2	Saturday	66.0	91.0	10.0	24.0	1.0	6.0	6.0
3	Sunday	73.0	94.0	11.0	25.0	2.0	7.0	7.0
4	Thursday	62.0	84.0	10.0	23.0	1.0	6.0	6.0
5	Tuesday	63.0	88.0	10.0	24.0	1.0	7.0	7.0
6	Wednesday	60.0	84.0	9.0	23.0	1.0	6.0	7.0

Fig 7: Day Wise Average Swipes/ Likes/ Messages - Sent/ Received

RQ5: Classification Algorithm (Machine Learning Algorithm)

- What is probability of having match for users knowing the features such as gender, swipes sent/ received, likes sent/ receive, messages sent/ receive etc., using machine Learning Algorithm. (**Classification Problem**)

Random Forest Algorithm is used to predict whether user will have a match with other user or not? This will help us in advance the potential candidates with having high probability of having match when become active. We can use this information for many purposes such as advertisements, promotion or any other purpose (like marketing of the application).

Classification Report – Evaluation

	precision	recall	f1-score	support
0	0.77	0.80	0.79	2915
1	0.83	0.80	0.82	3530
accuracy			0.80	6445
macro avg	0.80	0.80	0.80	6445
weighted avg	0.80	0.80	0.80	6445

Table 8: Performance Evaluation - Classification

Parameter	Value
Accuracy	80%
Precision	80%
Recall	80%
F1-Measure	80%

Classification Report – Evaluation

	Not Matched	Matched
Not Matched	2340	575
Matched	690	2840

The Algorithm trained well on the dataset.

Feature Importance

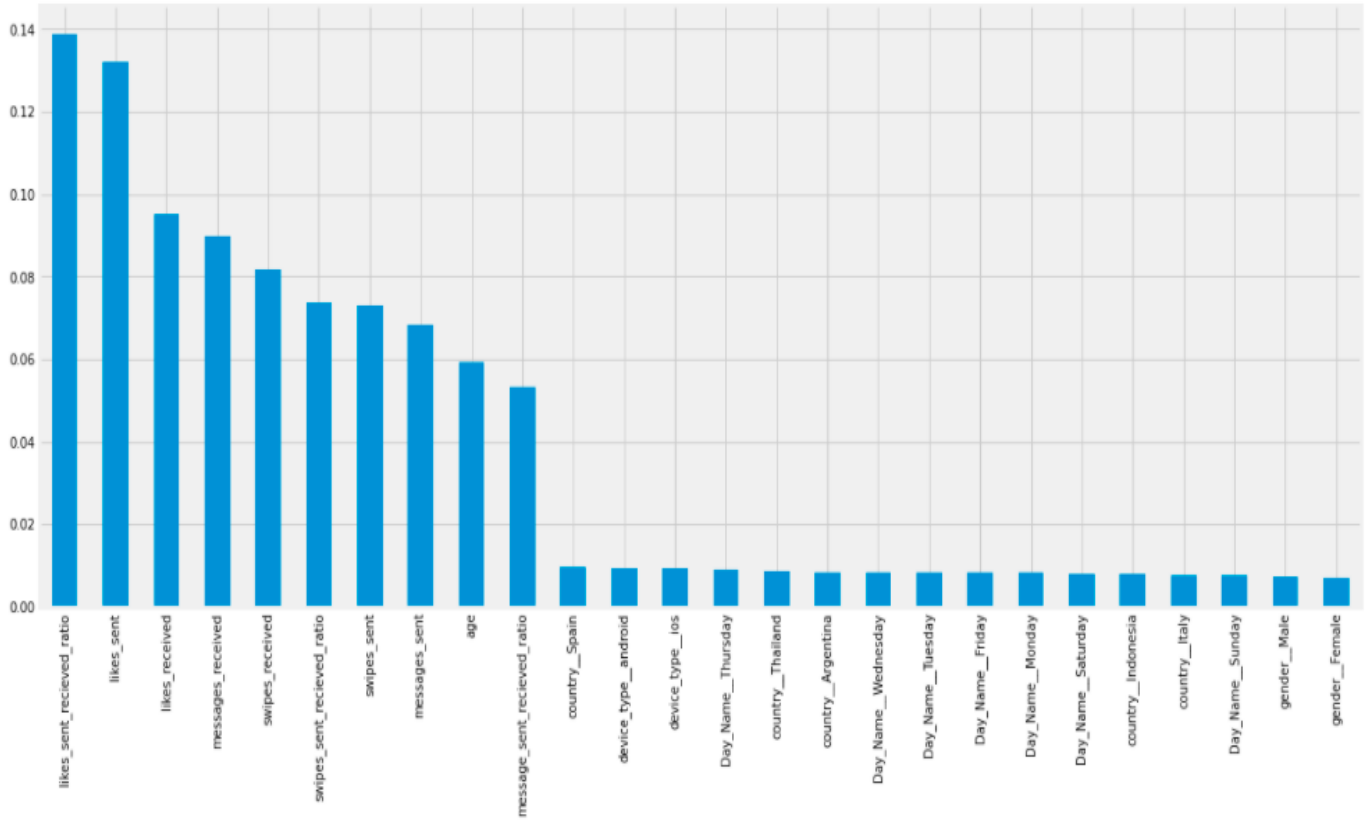

Fig 5: Feature Importance in Machine Learning - Classification

ROC Curve

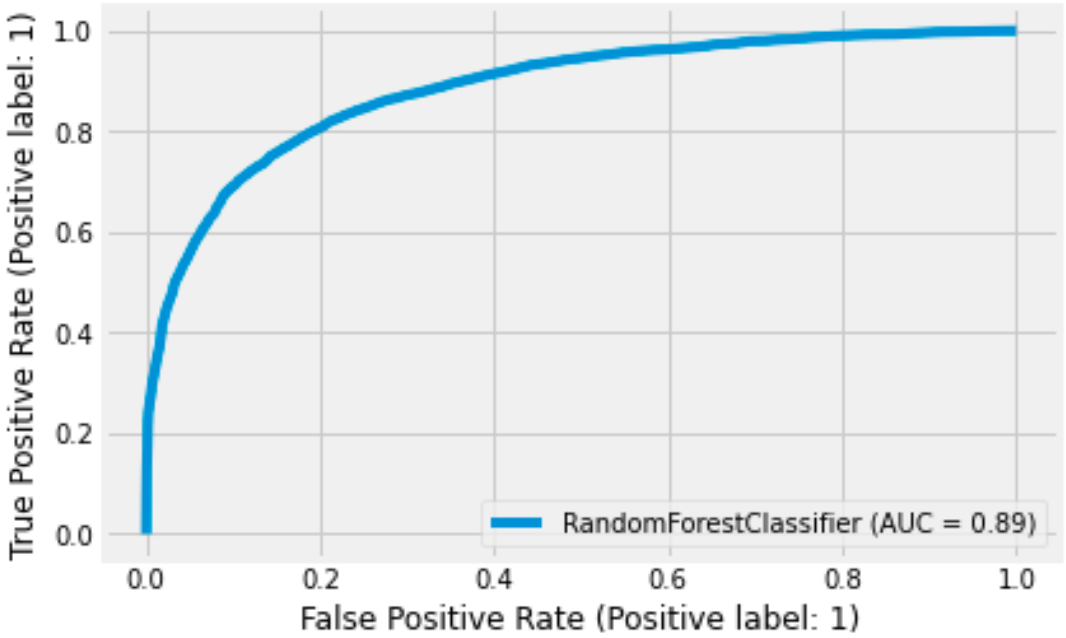

Fig 6: ROC Curve

RQ6: Regression Algorithm (Machine Learning Algorithm)

- Next time knowing above features, how many matches are possible in login using Machine Learning Algorithm (**Regression Problem**)

This method is used to find the number of possible matches when next time user become active on the basis of features such as swipes sent so far, received so far, likes or messages etc. Same as classification this will help application to increase customer interaction.

Evaluation Matrices

Table 9: Performance Evaluation – Regression

Parameter	Value
Mean Squared Error (MSE)	7.50
Mean Absolute Error (MAE)	1.22
Root Mean Square Error (RMSE)	2.73
R-Squared	0.74

Prediction for Evaluation of Algorithm

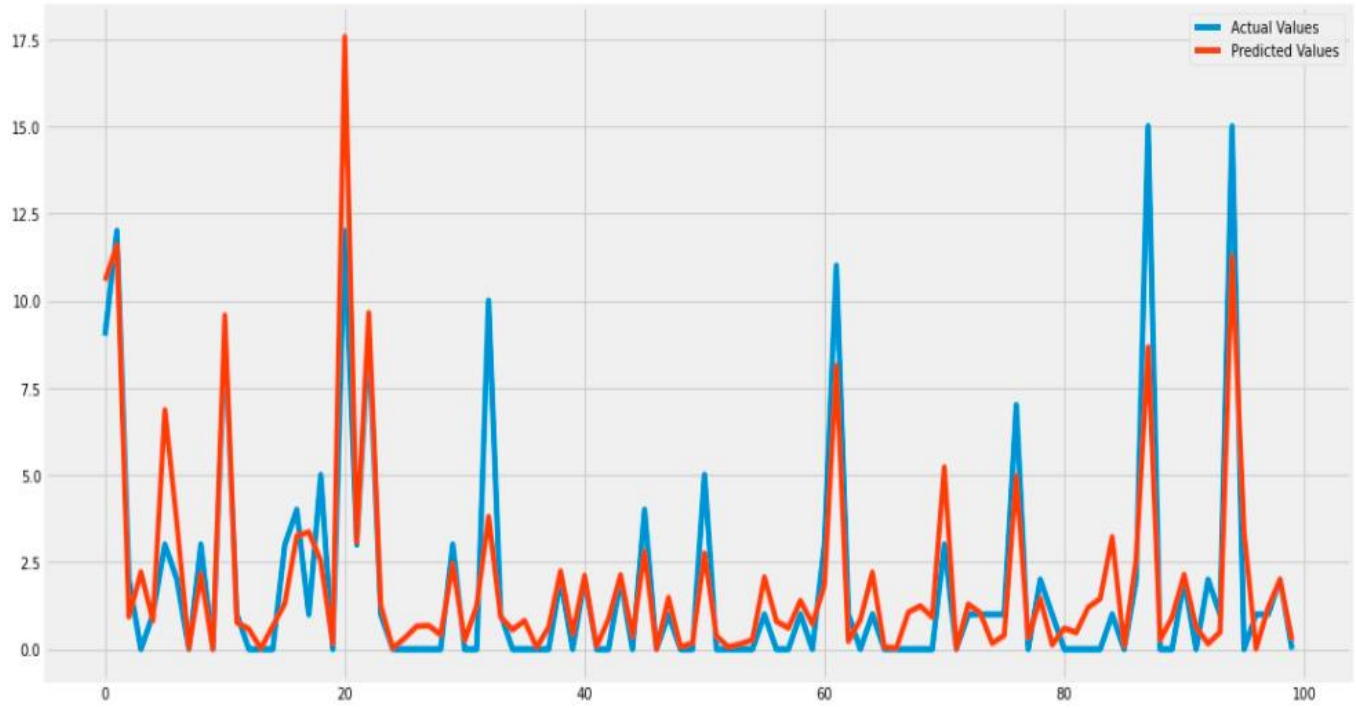

Fig 7: Actual Vs Predicted Values – Regression

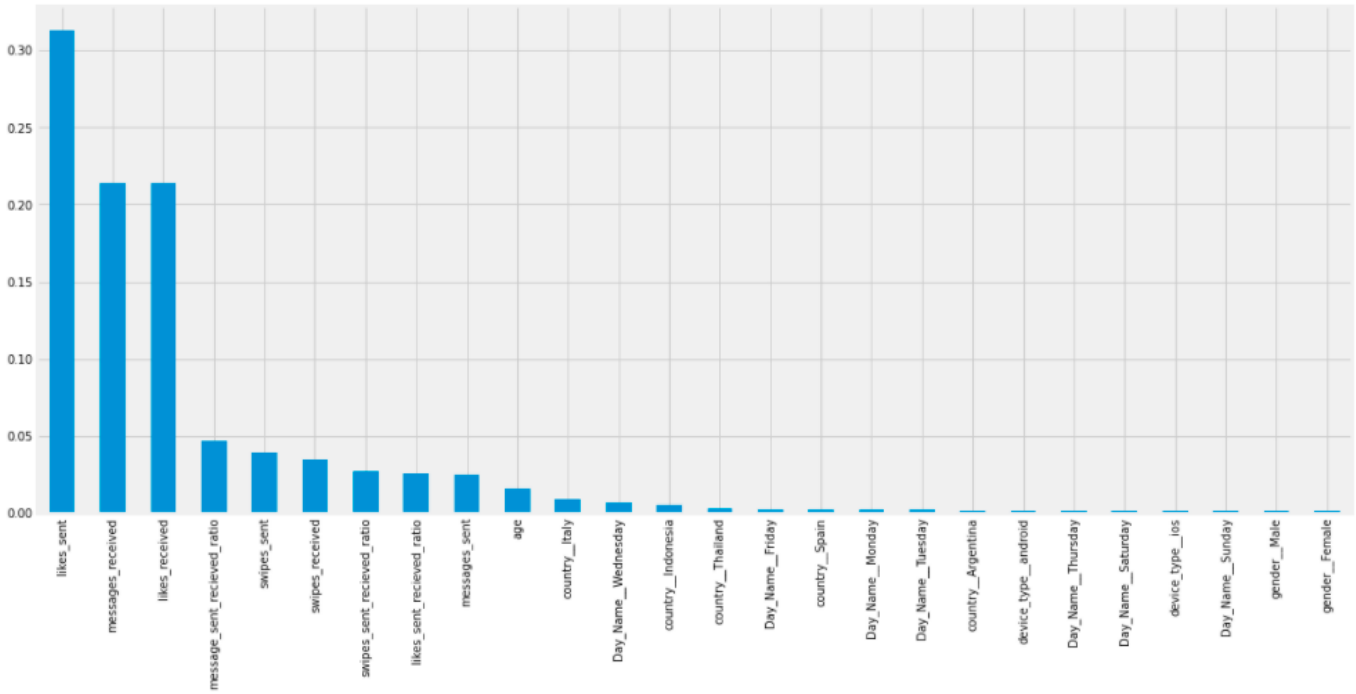

Fig 8: Feature Importance in Machine Learning – Regression

RQ7: Recommendation System

- Recommendation System: Recommending users which are most likely to pair-up/ match with each other based on available features such as likes, swipes and messages send/ received. (**using machine learning k-NN algorithm**)

This module will help us to increase user interaction with application. We can recommend users to user based on their similarities. This increase chance of user matches because similar users will interact with each other.

	2193
user_id	2194
age	19
country	Spain
gender	Female
device_type	android
is_active_user_average	1.0
swipes_sent_average	194.0
swipes_received_average	268.0
likes_sent_average	27.0
likes_received_average	192.0
matches_average	14.0
messages_sent_average	49.0
messages_received_average	61.0

	117	147	223	262	278
user_id	118	148	224	263	279
age	22	27	29	27	26
country	Italy	Italy	Italy	Italy	Italy
gender	Female	Female	Female	Female	Female
device_type	android	android	android	android	android
is_active_user_average	1.0	1.0	1.0	1.0	1.0
swipes_sent_average	25.0	88.0	27.0	207.0	591.0
swipes_received_average	167.0	14.0	19.0	538.0	289.0
likes_sent_average	1.0	14.0	10.0	3.0	14.0
likes_received_average	68.0	2.0	2.0	302.0	103.0
matches_average	0.0	1.0	0.0	1.0	2.0
messages_sent_average	1.0	0.0	6.0	20.0	26.0
messages_received_average	2.0	0.0	7.0	19.0	14.0

Fig 9: Recommended users to user ID: 2193

A recommendation on what you would do next with your work

The given dataset is useful for analysis, so I have performed basic exploratory data analysis as well some algorithms such as classification, regression and recommendation system. If you provide me full dataset (like the interaction between two users, other attributes of users, large data (of many months/ years), I can develop more sophiscated system. I can develop system which can improve system usage, we can increase users' interaction with system by identifying the users' behavior. I can write best match algorithm.