Find A Date: An Online Dating Recommendation System

ME781 Course Project Report, **Team 39**

Project Objective

- According to certain statistics, only around 60% of people find their online dating experience successful. The reason for this low satisfaction level, we believe, is the lack of emotional connect and common traits between people.
- Through this project, we aim to work with past online dating data and through detailed analysis, outline the major factors that lead to a successful dating experience. We also develop a ML powered recommendation system to improve dating experience.
- Additionally, we also deploy our system as an online web-app, allowing users to choose the best partners for them, supported by our convenient and speedy automated assistance :)

Problem Definition

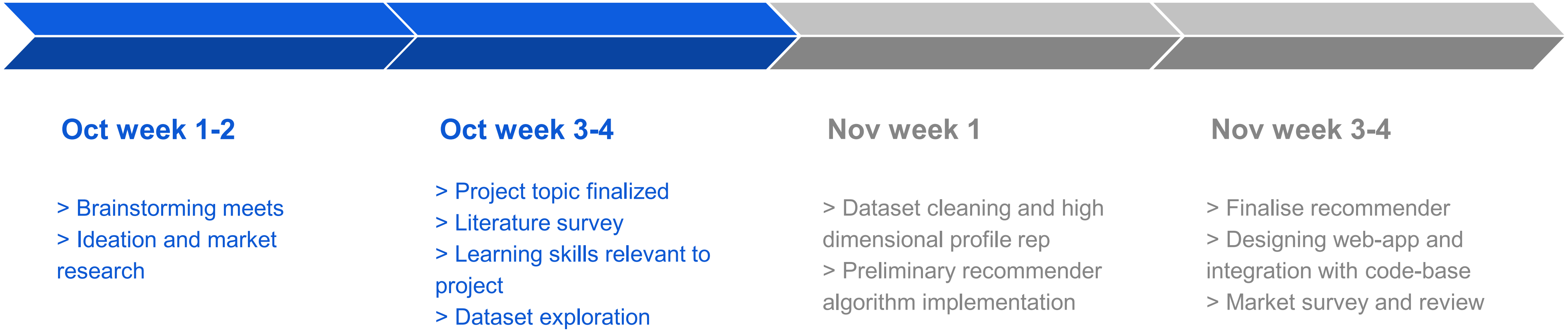
Customer Requirement	**An easy to use, convenient data-driven solution for filtered and personalised profile recommendations for online-dating**
Market Survey	• Tinder (primarily historical behavior and habit based ranking) • Bumble (mainly choice based filtering and activity based ranking) • OkCupid (math-driven calculation of compatibility score based on similarity of survey responses)
Key Differentiator	• Usage of latest machine learning and information retrieval algorithms to retrieve candidate profiles and rank them according to their predicted compatibility to the user • Usage of latent space representations, and reciprocal recommendation to augment ranking process
USP	• Proving latest tech-driven recommendation algorithm with an easy to use interface • Allowing the user to have control over various parameters of the recommendation algorithm, or default to the standard recommendation algorithm
Protect the USP	• Competent and powerful marketing coupled with a minimalistic and smooth user interface • Regular upgrades to the recommendation algorithm based on latest research and development • Encryption and reliable cloud services for security, copyrights and patents to protect the IP
Barrier to entry	• High advertising, legal and other costs to compete with giant online dating services • Lack of understanding among the common public about the technology being deployed

Technology Landscape Assessment

Literature	Helpful literature focused on reciprocal recommendation, collaborative filtering and regression based methods for scoring: - RECON: A Reciprocal Recommender for Online Dating - A Recommendation System Based on Regression Model... - Online dating recommendations ... - Design of reciprocal recommendation systems for online dating - Social Network Analysis and Mining
Open libraries	- **SciKit-Learn** has almost all functionalities needed to implement simple ML-based scoring algorithms, do dataset-preprocessing and evaluation - Other libraries such as Numpy, Pandas, Matplotlib, Scipy, etc. can be leveraged for data loading and processing, and to implement information retrieval and ranking algorithms
Dataset	- https://github.com/rudeboybert/JSE_OkCupid

Project Plan - Task Breakdown and Tentative Timeline

1. Choosing target features and recommendation workflow
2. Understanding the dataset, exploration and cleaning
3. Implementing representation of user profile pairs as feature vectors
4. Creation of a filtering pipeline to retrieve candidates for input profile
5. Application of various heuristic and ML algorithms for modelling profile similarity
6. Comparison of various ML and non-ML based recommendation methods
7. Development of the dating web-app and integration with codebase
8. Marketing Brochure and Marketing Video



Oct week 1-2

> Brainstorming meets
> Ideation and market research

Oct week 3-4

> Project topic finalized
> Literature survey
> Learning skills relevant to project
> Dataset exploration

Nov week 1

> Dataset cleaning and high dimensional profile rep
> Preliminary recommender algorithm implementation

Nov week 3-4

> Finalise recommender
> Designing web-app and integration with code-base
> Market survey and review

RASIC Chart - Roles and Responsibilities

Task	Gagan	Shubham	Sachin	Omkar
1	I	R	I	I
2	I	R	I	I
3	I	R	I	I
4	I	R	I	I
5	I	R	I	I
6	I	R	I	I
7	I	C, A	I	R
8	I	S, C, A	R	I

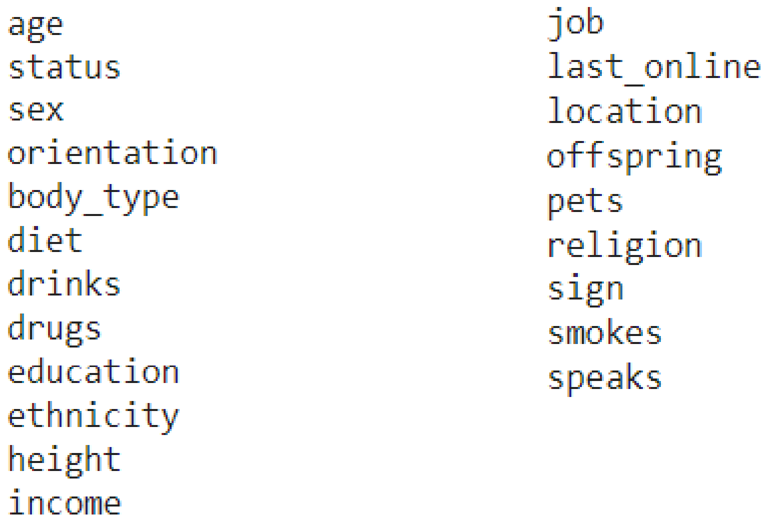
R - responsible, A - approve, S - supporting, I - informed, C - consulted

Project Design - Dataset

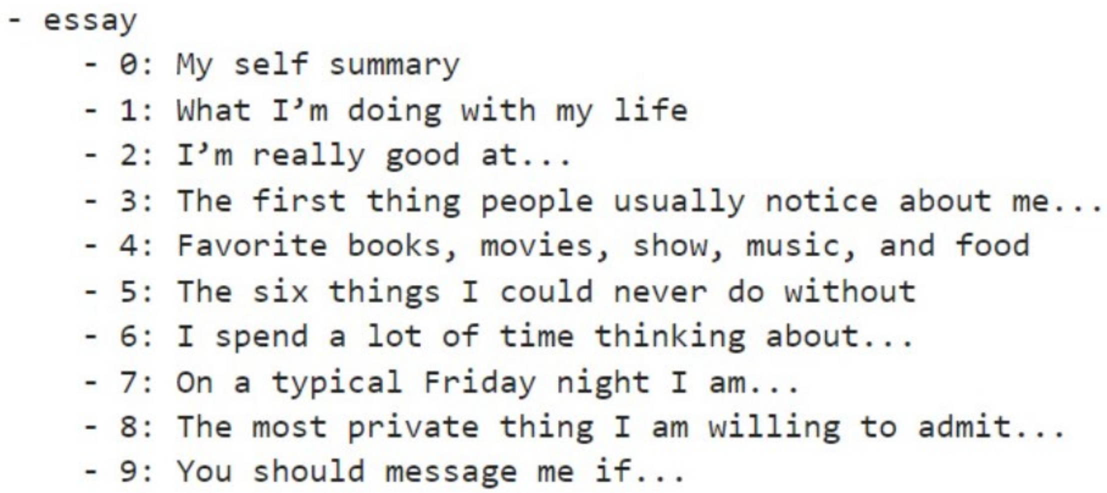
- OkCupid Profiles dataset from Kaggle has been used
  https://www.kaggle.com/andrewmvd/okcupid-profiles

- It has almost 60,000 online dating profiles

- The following attributes are present:



```
age              job
status           last_online
sex              location
orientation      offspring
body_type        pets
diet             religion
drinks           sign
drugs            smokes
education        speaks
ethnicity
height
income
```



```
- essay
    - 0: My self summary
    - 1: What I’m doing with my life
    - 2: I’m really good at...
    - 3: The first thing people usually notice about me...
    - 4: Favorite books, movies, show, music, and food
    - 5: The six things I could never do without
    - 6: I spend a lot of time thinking about...
    - 7: On a typical Friday night I am...
    - 8: The most private thing I am willing to admit...
    - 9: You should message me if...
```

Project Design – Data Filtering and Splits

- `'ethnicity', 'height', 'income', 'job', 'offspring’` were dropped because our aim is to make recommendations based on personality and personal choices
- `'speaks', 'last_online’` almost everyone spoke the same language, and last_online was redundant
- Only ‘single’ and ‘available’ status people were kept, rest were deleted from recommendation database
- Profiles in locations with less than 5 profiles were dropped, as we plan to return topK profiles, so in locations with very few profiles, personalised recommendation doesn’t make sense
- Data was split into a train-test split of 60:40. The training data helps train the recommendation model, and as the database of profiles. The testing data serves as query data for which we make recommendations from profiles in the training data

Project Design – Train Data Generation

- Feature Functions are defined for each attribute to return a score from 0 to 1 based on heuristical methods which capture how similar the attribute values are to each other. The vector of all such attribute similarity scores makes up the compatibility vector of a profile pair

- We reserve some attributes (‘pet’ preference, ‘smoke’ preference, and ‘About me’ essay) for proxy labelling to introduce supervision signals in this semi-supervised learning approach

- For each profile in the training data, we filter the other available profiles based on location and sexual orientation compatibility, and then sample a maximum of 10 profiles, and add the compatibility vectors and proxy-labels of these pairs to the training data for our supervised learning models

Detailed description of the feature compatibility score generation, proxy label assignment, and training data generation process can be found in the detailed code documentation.

Project Design – Recommendation Generation

- After the models have been trained, saved checkpoints are loaded and can be used to get recommendations
- For a query profile, first the available profiles are filtered based on location and sexual orientation compatibility to get the candidate set, and then probability scores of recommendation are taken from the classification model output
- These scores are then ranked in descending order, and the top ‘K’ profiles are returned as recommendations. We set K to 5. Also as the dataset is too big, we run the model on randomly sampled 100 profiles from the candidate set

Project Design – Evaluation

- To evaluate the recommendation model, we do the following:
  - If probability score > threshold (hyperparameter), we assign label as 1
  - We check how many profiles assigned label 1 also had proxy-label assignment as 1. The profiles that match are termed relevant profiles
  - If fraction of relevant profiles in top K > 0.6 (hyperparameter), we define the recommendation as successful
  - The fraction of successful recommendations on a test queries dataset is termed as the relevancy score of evaluation

- Model Selection is done on the basis of relevancy score.
- High relevancy score on high thresholds implies a better recommendation model

Project Design – Evaluation Report

Test set performance

		Model		
Threshold		Logistic Regression	Multi-Layered Perceptron	Naïve model (score = feaure average)
	0.25	0.7956	**0.7994**	0.7862
	0.5	0.7845	**0.7858**	**0.0074**

- MLP is the best model for profile recommendation
- Note that the naïve scoring model also supposedly does well at low thresholds, but when relevancy threshold is increased, it fails miserably, but our supervised learning models hold up with very little decay in performance. This demonstrates our framework's capability to provide recommendations

User Manual – Recommendation Framework Codebase

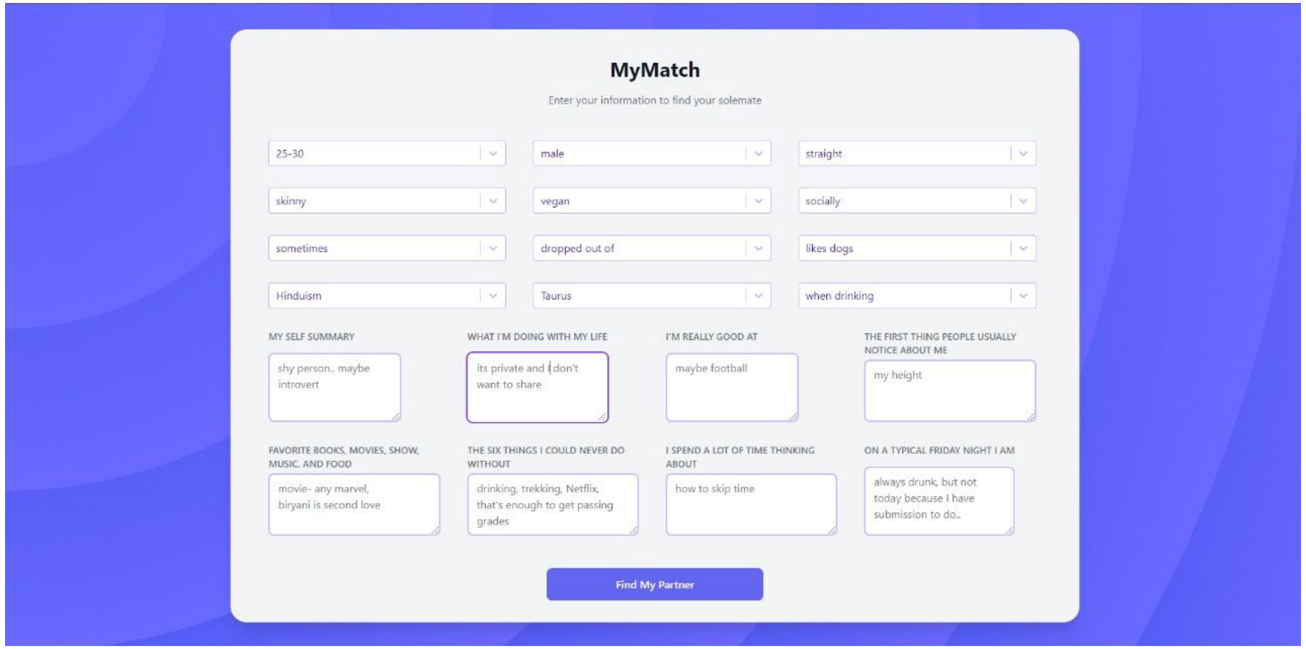
- The code of the project has been extensively documented, and the documentation can be accessed at https://shubhlohiya.github.io/dating-profile-recommendation/

- Detailed Run Instructions are available at https://github.com/shubhlohiya/dating-profile-recommendation/blob/master/README.md

User Manual – User Interface

1) Fill form with your preferences 2) Wait for recommendations 3) Browse recommendations

(This web-app frontend was build using React. A python script runs on the backend to return recommendations.



MyMatch

Enter your information to find your solemate

25-30	male	straight
skinny	vegan	socially
sometimes	dropped out of	likes dogs
Hinduism	Taurus	when drinking

MY SELF SUMMARY: shy person.. maybe introvert

WHAT I'M DOING WITH MY LIFE: its private and I don't want to share

I'M REALLY GOOD AT: maybe football

THE FIRST THING PEOPLE USUALLY NOTICE ABOUT ME: my height

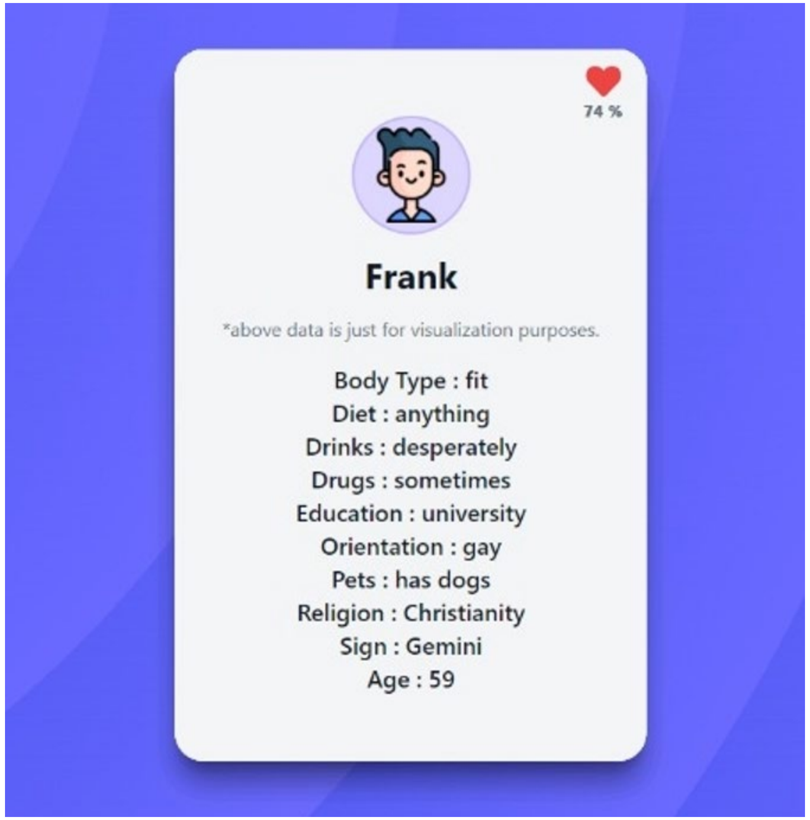
FAVORITE BOOKS, MOVIES, SHOW, MUSIC, AND FOOD: movie- any marvel, biryani is second love

THE SIX THINGS I COULD NEVER DO WITHOUT: drinking, trekking, Netflix, that's enough to get passing grades

I SPEND A LOT OF TIME THINKING ABOUT: how to skip time

ON A TYPICAL FRIDAY NIGHT I AM: always drunk, but not today because I have submission to do..

Profile Form



Frank

*above data is just for visualization purposes.

Body Type : fit
Diet : anything
Drinks : desperately
Drugs : sometimes
Education : university
Orientation : gay
Pets : has dogs
Religion : Christianity
Sign : Gemini
Age : 59

Recommendation Card