IS597MLC-SP24: Final Project Proposal

NetID: houchun2
Student Name: Jimmy Liu

Title

Zodiac Prediction via Dating Profiles: A Data-driven Approach

Motivation & Objective

Background and Motivation:
Astrological signs, also known as zodiac signs, are believed by many to influence an individual's personality traits, behavior, and compatibility with others. This belief has persisted for centuries, capturing the imagination of people across cultures. In the realm of dating and relationships, the notion of zodiac compatibility is widely discussed and considered by some as a factor in determining potential romantic matches. However, the validity of these astrological claims remains a subject of debate, with skeptics dismissing them as pseudoscience. In this project, I aim to take a data-driven approach to investigate the rumored connection between zodiac signs and personal characteristics, as reflected in dating profiles.

Objective:
The primary objective of this project is to develop a machine learning model that can accurately predict an individual's zodiac sign based on their dating profile data. By analyzing the patterns and relationships between profile attributes (such as interests, personality traits, and preferences) and zodiac signs, I hope to gain insights into whether there is indeed a correlation between astrological signs and specific personality or behavioral tendencies.

Research Questions:
1. Can machine learning algorithms effectively predict an individual's zodiac sign based on their dating profile data?
2. Which features or attributes within dating profiles are most predictive of an individual's zodiac sign?
3. If a correlation between zodiac signs and personal characteristics is found, how strong is this relationship, and what are the potential implications for understanding human behavior and compatibility?

By addressing these research questions, I aim to contribute to the ongoing discourse surrounding the validity of astrological beliefs and their relevance in the context of dating and relationships. Moreover, this project has the potential to shed light on the predictive power of machine learning techniques in areas often considered unconventional or controversial.

Related Articles

The first article (van der Lee, van der Zanden, Krahmer, Mos, & Schouten, 2019) compares the performance of the lexicon-based text analysis tool LIWC, machine learning models, and human judgments for predicting writers' intended relationship goals (long-term vs. casual dating) from their online dating profile texts. Using a new corpus of Dutch dating profiles, the study evaluated these methods of assigning relevant psychological labels and accurately classifying the stated relationship goal. The results showed that LIWC's labels aligned more closely with human judgments for some categories. Still, all three approaches performed similarly in predicting the relationship goal itself, while LIWC's fixed lexicon had some limitations for this text genre.

The second article (He et al., 2021) proposes DatingSec, a novel system for detecting malicious accounts in dating apps. DatingSec utilizes a content-based attention network to analyze users' static features (profile, community) and dynamic features (behavior, posts, comments). The key novelty is incorporating textual information from user interactions and using an attentive module to capture suspicious patterns. Evaluation on a real-world dataset from the dating app Momo shows DatingSec outperforms existing methods.

A recent study has explored sentiment analysis, machine learning, and user profiling in online social environments, notably in predicting gender based on social media content. A study integrated sentimental features into existing machine learning frameworks through sentiment analysis and transfer learning techniques (Liu, Li, & Li, 2021), demonstrating its efficacy in discerning emotional tones and linguistic patterns indicative of gender identity.

To find the scientific evidence that zodiac signs influence a person’s characteristics, we look at the fourth article. The article (Helgertz & Scott, 2020) examines the validity of astrological predictions on marriage and divorce by analyzing longitudinal data from Swedish registers over the period 1968-2001. It tests various classifications of astrological compatibility between partners based on their zodiac signs, looking at the prevalence of such pairings among married couples and their risk of divorce. The results fail to provide consistent evidence supporting the notion that astrologically more compatible couples are overrepresented among marital unions or have a lower risk of divorce.

Data

A. Data Collection

The dataset (Kim & Escobedo-Land, 2015) is derived from the dating app OkCupid. It comprises public profiles of 59,946 OkCupid users who resided within 25 miles of San Francisco, had active profiles during a period in the 2010s, were online in the previous year, and had at least one picture in their profile. Each profile contains 31 attributes. The data was scraped from users' public profiles using a Python script, excluding non-publicly facing information such as messaging. The variables include typical user information (e.g., sex, sexual orientation, age, ethnicity) and lifestyle variables (e.g., diet, drinking habits, smoking habits). Random noise was added to specific variables for de-identification purposes. The columns "Essay 0~9" and "Sign" are the focus of the analysis.

Attribute Name	Description
Age	The age of the user
Status	Single or seeing someone
Sex	Male or femail
Orientation	Straight, Gay, Bisexual
Body Type	Average, fit, thin… etc
Diet	Diet type of the user
Drinks	Drinking frequency of the user
Drugs	Drug using habit of the user
Education	Education level of the user
Ethnicity	White, Asian…etc
Height	User height in inches
Job	User job
Last_online	Last online datetime
Location	User location (mostly bay area)
Offspring	Number of offspring
Pets	Like dogs or cats
Religion	Religion of the user
Sign	Zodiac sign of the user
Smokes	Smoking habit of the user
Speaks	Language the user speaks
Essay0	My self summary
Essay1	What I’m doing with my life

Essay2	I’m really good at
Essay3	The first thing people usually notice about me
Essay4	Favorite books, movies, show, music, and food
Essay5	The six things I could never do without
Essay6	I spend a lot of time thinking about
Essay7	On a typical Friday night I am
Essay8	The most private thing I am willing to admit
Essay9	You should message me if...

B. Data Pre-processing

In the data cleaning process, the first step is to handle missing values. The text mentions that rows with missing values in the "Sign" category will be dropped. As for missing essay responses, those will be filled with blank strings.

After addressing missing values, the next step is to perform text preprocessing techniques. This includes:

1. Stripping the essays to remove leading/trailing whitespace characters.
2. Tokenizing the text in the essays using the Natural Language Toolkit (NLTK) library. Tokenization is the process of breaking down text into smaller units called tokens (words, phrases, etc.).
3. Identifying and removing stopwords (common words like "the", "a", "and", etc.) from the essays using NLTK or other NLP libraries. This helps create a higher quality dataset by removing words that don't contribute much meaning.
4. Applying stemming techniques using NLTK or other NLP libraries. Stemming is the process of reducing words to their root/stem form (e.g., "learning", "learned", "learns" -> "learn").

After these steps, the dataset will be ready for vectorization, which converts the text data into numerical form that can be used by machine learning models. Regarding the label/target class, the text mentions using the "Sign" category as the label. Since the model can directly use categorical data as input, the "Sign" column may be kept in string format for better recognition by the model.

To summarize, the data cleaning process involves handling missing values, text preprocessing (stripping, tokenization, stopwords removal, stemming), and preparing the label column. Libraries like NLTK, along with Python's built-in string manipulation functions, will be used for these tasks.

Analysis & Methodology

To achieve my goal, I plan to use the following techniques and methods:

I would use all the essays as input and employ the TF-IDF Vectorizer to convert the textual data into numerical feature vectors. A linear classifier, such as Logistic Regression or Support Vector Machines (SVM), will be used as the base model to establish a baseline performance.

After building the baseline model, I would start conducting the following experiments:

1. Test if different essays have different influences on the result. Perform feature selection and engineering for the input essays. Explore keyword extraction and a combination of different essays to find the best input features.
2. Adjust the vectorizer to more advanced embedding models, such as GloVe and BERT.
3. Try out different models like Random Forest, SVM, etc. After trying these, I would attempt to build a deep learning model with an attention mechanism using PyTorch.

For evaluation metrics, I will initially use accuracy for basic evaluation. After achieving a stable accuracy, I would examine the precision, recall, and F1 score for each class. Since there are 12 different zodiac signs, I would analyze the balance of each class and determine if class-balancing techniques need to be employed for training and testing.

References

Kim, A. (n.d.). Rudeboybert/JSE_OKCUPID: Journal of Statistical Education Paper on using okcupid data for data science courses. GitHub. https://github.com/rudeboybert/JSE_OkCupid?tab=readme-ov-file

Albert Y. Kim & Adriana Escobedo-Land (2015) OkCupid Data for Introductory Statistics and Data Science Courses, Journal of Statistics Education, 23:2, DOI: 10.1080/10691898.2015.11889737

He, X., Gong, Q., Chen, Y., Zhang, Y., Wang, X., & Fu, X. (2021). DatingSec: Detecting malicious accounts in dating apps using a content-based attention network. IEEE Transactions on Dependable and Secure Computing, 18(5), 2193-2205. https://doi.org/10.1109/TDSC.2021.3068307

van der Lee, C., van der Zanden, T., Krahmer, E., Mos, M., & Schouten, A. (2019). Automatic identification of writers' intentions: Comparing different methods for predicting relationship goals in online dating profile texts. Proceedings of the 5th Workshop on Noisy User-Generated Text (W-NUT 2019). https://doi.org/10.18653/v1/d19-5512

C. Liu, F. Li and L. Li, "Research on Gender Prediction for Social Media User Profiling by machine learning method," 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), Beijing, China, 2021, pp. 831-836, doi: 10.1109/CISCE52179.2021.9445922.

Helgertz, J., & Scott, K. (2020). The validity of astrological predictions on marriage and divorce: a longitudinal analysis of Swedish register data. Genus, 76(1), 1-18. https://doi.org/10.1186/s41118-020-00103-5