**Data Project: Dating Analysis**

Reed Foster, Jack Sheridan, Joseph Kim, Patrick Foster, Gary Lee

The University of Iowa

BAIS:3500

12/9/2023

**Executive Summary**

*The following report outlines the underlying problem and analysis done by us at MatchMaker Analytics to address Tinder’s urge to improve user connections. In response to this imperative, we have undertaken a comprehensive analysis of historical dating data from 2002-2004 to reveal tendencies related to this issue. Focusing on key attributes such as attractiveness, intelligence, fun, ambition, and shared interests, the team streamlined a dataset from Kaggle to 16 critical features. Through exploratory analysis and data preprocessing, the study revealed insights into user preferences, particularly highlighting the significance of intelligence and ambition in partner selection. Leveraging machine learning, a logistic regression model was developed and fine-tuned, achieving an impressive 85% accuracy. With a focus on minimizing false positives, the model demonstrates robust performance, paving the way for Tinder to implement data-driven strategies that could significantly improve the success rate and overall user experience on their platform. Our overall recommendation is to implement our model, in which benefits could then be built out in the actual physical interface as well as the suggestive algorithm. Although this is a great starting point to continue to test the model, there are several limitations in our data set such as the time the data was collected.*

**Background**

In the fast-paced world of online dating, making meaningful connections can be a challenging endeavor. **Tinder**, a prominent online dating platform, has identified a pressing business challenge: optimizing the success rate of their matches. Tinder believes that by delving into data from the past, they can unravel the underlying factors that lead to fruitful encounters. This endeavor falls under the purview of the Data Insights Team at **MatchMaker Analytics**, a trusted partner in data-driven solutions. The mission is to formulate a strategy to achieve Tinder’s aspirations of enhancing the success rate of their matches by analyzing data about the characteristics of past successful matches. Our analysis will look at unique combinations of characteristics from successful pairings in the past so that we can alter algorithms to cross the paths of these combos. The specific dataset, sourced from Kaggle, captures participant feedback from experimental dating events held between 2002-2004, focusing on attributes such as attractiveness, sincerity, intelligence, fun, ambition, and shared interests. This process will set the foundation for uncovering crucial insights that will pave the way towards more successful and fulfilling dating experiences for Tinder’s users.

**Business Goal:**

The overarching business goal is to enhance the success rate of matches on Tinder. By leveraging data analytics, we are aiming to refine Tinder’s match making algorithms, ultimately leading to more fulfilling and successful dating experiences for its users. MatchMaker Analytics is tasked with the mission of formulating a data-driven strategy to achieve this objective.

**Data Mining Goal:**

The data mining problem revolves around predicting the likelihood of two individuals being a match based on various specified factors. The prediction target is a binary outcome – whether or not a match will occur.

The specific data mining goals include:

* Developing and fine-tuning predictive models to determine match outcomes.
* Uncovering attribute pairings that significantly contribute to successful matches
* Addressing potential challenges through rigorous model evaluation and experimentation with different classification algorithms.

Overall, the data mining goals align with the broader business objective of improving the success rate of matches on Tinder by leveraging insights derived from historical dating data.

**Data Description**

The raw dataset we are using is from Kaggle. It was collected from experimental dating events during the years 2002 – 2004. The participants in these experiments were asked to give insight on their date such as a rating within six attributes: Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests. They also shared personal information such as dating habits, what they believe is valuable, and demographic information. As mentioned previously, the raw data set came with 122 features, but we reduced it to the 16 most important features.

**Exploratory Analysis: (**[**EA #1**](#Bookmark3) **/** [**EA #2**](#Bookmark4)**)**

Our team looked at a series of graphs and ties within the data to get a better understanding of our data set and how it could really help with Tinder’s problem. Below is a series of those analyzations and our findings.

A diagram of value of different sizes and colors

Description automatically generated with medium confidence

*We found that those who rate themselves as being extremely funny have a much lower chance of a match compared to their counterparts. Additionally, there is an inverse correlation between a partner looking for sincerity and attractiveness.*

A grid of blue and green dots

Description automatically generated

We wanted to understand if there's a connection between the age difference among speed dating participants and the attractiveness rating given by their partners. This could help app users determine whether age determines perceived attractiveness. that, with gender represented in the plot, we can explore how ratings differ between genders. This can be interesting if we want to examine whether gender plays a role in how attractiveness is rated.

A graph of red and blue dots

Description automatically generated

We wanted to see visual evidence of the correlation between features that are tied together like funnny\_o\_ranking and funny\_partner to ensure that the model we are shooting for is realistic. As you can see there is a strong correlation between a person’s personal funny ranking and whether they got a funny partner.

A screenshot of a computer

Description automatically generated

We also wanted to explore the strength of various relationships in the data set. The correlation widget helped us identify and understand correlations that were useful as we moved into feature selection. By analyzing these numbers, we could avoid overfitting, pinpoint data quality issues, and identify trends.

A screenshot of a computer

Description automatically generated

We wanted to investigate the importance of race when choosing a partner in this dataset to get a better understanding of what features are significant. 1 is considered a match and 0 is not a match. As it is shown here race is not a large determining factor whether or not there is a match. The mean ranking out of 10 for both match and nonmatch is not that different signaling it is not a significant factor. The overall rating is also low, meaning that most participants do not see it as an important factor.

A screenshot of a computer

Description automatically generated

We wanted to explore how important attractiveness was for the potential of a match. This bar graph shows the amount of people that ranked attractiveness importance on a scale of 1 to 10 for each number. As it is shown most people rated attractiveness as an important factor since most rankings are between 5 and 9. This could be a good feature to help predict a final match.

A graph of a bar chart

Description automatically generated with medium confidence

We also wanted to look at how important intelligence is when searching for a partner. After creating the EDA, we found out that intelligence is very important for males and females. With both genders the results are skewed left with majority of the results ending up between 6-10. For men, the main range of importance is 5-8, while for females it is 8-10. This means the men hold intelligence to a medium high importance while females hold intelligence to be one of their most important factors in choosing a partner.

**Data Pre-processing**

For our data preprocessing we thought that 122 features were too many to use for conclusive results, so we narrowed them to the most important 16 features. To get down to the main 16 we had to take out certain features and combine certain ones using orange to create 16 features that would accurately describe our data. The 16 key features that were used can be found in the [data dictionary](#Bookmark5). Each person going on the date specified how much they valued attractiveness, sincerity, intelligence, humor, and ambition in a partner. Then for each speed date, they rated each person on these attributes. We did feature engineering for each of these attributes, for both people going on the date, where we measured the discrepancy between desired and actual for each attribute.

The equation was for each attribute: `1 - |(expected / max-expected) - (actual / max-actual) |`

For example, if they valued humor at 75 (on a scale of 100) and rated then person as a 3 (on a scale of 10), this feature would be `1 - |(75/100) - (3/10) |` = 0.55, indicating a mild match.

We imputed missing values as averages and standardized continuous variables.

**Data Mining Solution:**

To help Tinder match individuals that will go on dates, we built a machine learning model to predict if two individuals are likely to be willing to go on a date with one another. The model should predict a zero if they are unlikely to go on a date and 1 if they are likely to go on a date. Tinder can later implement this model into their app’s feed of suggested partners.

When developing the model, we had the cost for False Positives (predicting they should go on a date) be much higher than a False Negative (predicting they should not go on a date). Our rationale was that a False Positives would leave a poor taste in our customers' mouths, having been matched with someone they lack a connection with. Conversely a False Negative would merely hide an individual from seeing someone else, having little impact to either party.

We tested numerous different models, all performing classification: Neural Networks, Gradient Boosting, Decision Trees, and Logistic Regression. However, the models consistently performing the best were Logistic Regression. After tuning the hyperparameters we reach a model with an accuracy of 85%. It used a Regularization type of *Lasso* and *C* equaled 0.11 for the model’s strength.

*Performance metrics:*

*AUC: 85%*

*CA: 86%*

*F1; 83%*

*Precision: 85%*

*Recall: 86%*

*Out of 2513 predictions, there were only 42 false positives.*

**Recommendations**

Our team has come up with a suggestive plan that you could implement to test these findings. One way you could implement and test the results provided by our model is by selecting a target market, say Brazil for example, and add functionality for these users that requires them to fill out a survey that includes questions tied to the attributes within our data set like the level to which a user prioritizes intelligence. Once you have this data you have multiple options to utilize it for better connections. First, those preferences could then be publicly visible to other users so they can see and make decisions based off that, or you could also implement the factors within the data set as identifiers in your algorithm, so people with preference similarities are more likely to see each other. Like we stated before you could roll this out in Brazil and test the results in comparison to other areas. We touched on earlier, there may be other important attributes to consider. Regardless, we think this is a great starting point in which you could continue to test and add to the model.

**Limitations**

Although we recommend the implementation of this model, there are several limitations that are important to take into consideration. First, there are demographic limitations. The data in our data set was collected between 2002 and 2004. With it now being 2023, this data is rather old and may not be entirely accurate in terms of accurately representing the tinder user today. Second, there are locational limitations. The data set was focused on specific regions, so the yielded results may not be fully representative of Tinder’s diverse global use base. Third, the data set may be too generalized. There may be factors that are missing from the data set that could also be beneficial to use for analysis. Lastly, imputing and filling in missing values in the dataset can lead this model to be unequipped to handle all the broad use cases Tinder has. These four limitations are important to keep in mind as you work towards the implementation of this model.

**Future Work**

MatchMaker Analytics' work is a first step in a journey for Tinder transforming their match making algorithm. It will take a dedicated team of software engineers to deploy the model we created and integrate it into Tinder’s current algorithm. We expect a cloud hosting provider such as AWS or Microsoft Azure will be necessary in this deployment. Furthermore, as previously recommended, this will be an iterative implementation. Operating a machine learning model at the scale of Tinder is computationally intensive, requiring sophisticated cloud computing with a gradual implementation. Furthermore, as data is collected on matches at Tinder, a more accurate model could be trained.

**Data dictionary:**

|  |  |  |
| --- | --- | --- |
| Column | Data Type | Description |
| age\_o | Numeric | Age of partner |
| race | Numeric | Race of self |
| race\_o | Numeric | Race of partner |
| importance\_same\_race | Numeric | How important is it that partner is of same race? |
| pref\_o\_attractive | Numeric | How important does partner rate attractiveness |
| pref\_o\_sincere | Numeric | How important does partner rate sincerity |
| pref\_o\_intelligence | Numeric | How important does partner rate intelligence |
| pref\_o\_funny | Numeric | How important does partner rate being funny |
| pref\_o\_ambitious | Numeric | How important does partner rate ambition |
| attractive | Numeric | Rate partner- attractiveness |
| sincere | Numeric | Rate partner - sincerity |
| intelligence | Numeric | Rate partner - intelligence |
| funny | Numeric | Rate partner - being funny |
| ambition | Numeric | Rate partner - ambition |
| Attractive match | Numeric | 0 – 1 score on how well attractive matched between actual and expectations |
| Sincerity match | Numeric | 0 – 1 score on how well sincerity matched between actual and expectations |
| Intelligence Match | Numeric | 0 – 1 score on how well intelligence matched between actual and expectations |
| Funny Match | Numeric | 0 – 1 score on how well being funny matched between actual and expectations |
| Ambition Match | Numeric | 0 – 1 score on how well ambition matched between actual and expectations |
| match | Categorical | Match (yes/no) |

[Data Description](#Bookmark2)