Application of Machine Learning to Predicting Matches in Speed Dating

Introduction

Problem statement

Recent speed dating events have not been particularly successful, with many participants leaving without a match. Participants frequently mention that they spend most of the evening talking to people that they don't feel any connection with, and as a result leave feeling dissatisfied with the event. This customer dissatisfaction has led to poor reviews and fewer word of mouth recommendations, which is impacting the business's reputation, demand and ultimately revenue.

Proposed solution

A potential solution to this problem would be to pair participants up with partners that they are more likely to match with. This could be done for example by taking details from a large pool of participants, identifying their likely matches, and then assigning them to smaller speed dating events where the participants are selected based on their compatibility. The advantage of this approach is that participants will spend less time talking to people that they don't feel a connection with, and are more likely to leave the event with a match.

Project objective

The aim of this project is to investigate what factors affect the likelihood of a match, and to use supervised machine learning techniques to develop a predictive model that aims to predict whether or not two participants will be a match before they've met. This model can then be used to pair up participants with their likely matches, thereby increasing customer satisfaction, business reputation and revenue.

Analysis Plan

The plan for the project is as follows:

- Import and clean the data, readying it for investigation
- Perform exploratory data analysis to extract relevant insights
- Select appropriate performance metrics for use in evaluating a machine learning model for predicting matches
- Identify and fine-tune a suitable model to maximise its predictive power
- Evaluate the results to determine if the model is ready for use or if further work is required

The Dataset

The dataset used in this project can be found on DataCamp's GitHub account (https://github.com/datacamp/careerhub-data/tree/master/Speed%20Dating), as well as details of the column names and their descriptions.

Each row of the dataset contains information about the person who is going on a date, including age, race, interests, self-ratings on various qualities, and ratings of the importance of said qualities in a partner. From now on we will refer to this individual as Person A. Additionally there is some information about the other individual attending the date, though this data is much more limited; for example we have no information on their interests or self-ratings. We will refer to this individual as Person B.

Cleaning the Data

Streamlining the dataset

The first step was to drop columns that are not useful for this analysis. There are many variables in the dataset that can only be obtained after the date, such as 'sincere_o' (sincerity rating of Person A by Person B) and 'attractive_partner' (attractiveness rating of Person B by Person A). These variables were dropped because we are only interested in predicting matches based on information we can obtain ahead of the date. Additionally there are a number of variables beginning with d_. Aside from d_age, which is the age difference between the individuals, these columns simply group other numerical features into broad buckets. These variables were also dropped as they don't add any new information to the dataset. Finally any further columns that were unlikely to contribute to the model were dropped; in particular 'has_null' (which just indicates whether there are any null entries in the row) and 'field' (which has too many possible answers to be encoded as a categorical variable).

Removing duplicates and missing values

The next step was to check for duplicates and missing values. While no duplicates were found, there were a number of missing values. In particular, two columns had over 1,000 missing values; 'expected_num_interested_in_me' and 'expected_num_matches'. Given the large proportion of missing values in these columns, it was decided to drop them entirely, rather than to guess what the values should be and potentially bias the predictive model. Further analysis showed that the remaining missing values related to a very narrow selection of rows, comprising just 3.8% of the dataset. As this is such a small proportion of the dataset, and keeping the rows would have required guessing values across many columns with no information to base it on, these rows were dropped from the dataset.

Checking data types and outliers

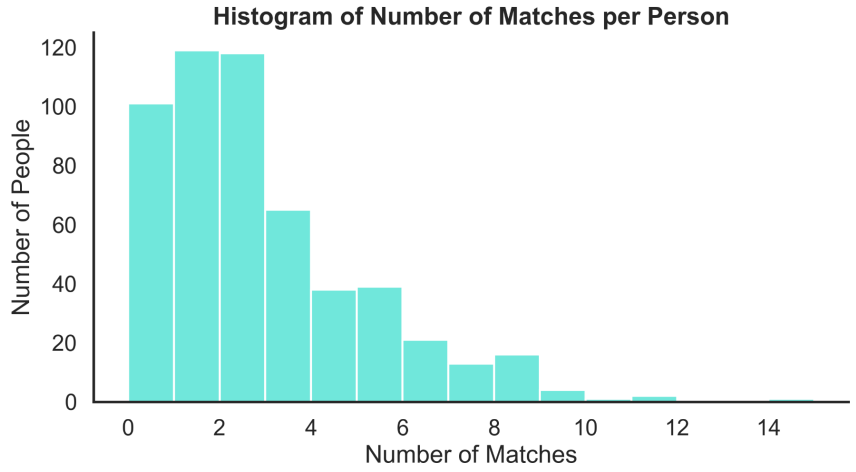
The next steps involved checking the data types of each field and ensuring that the data sat within the expected parameters for each field. Some columns had inappropriate data types - typically those that should be integers or boolean values were encoded as floats, and categorical variables were encoded as objects - so these were corrected. There were also some values that sat outside of the expected range - for example, there were some values of 13 and 14 in fields that should have been between 0 and 10. On the assumption that these were typos and were intended to be 3 and 4 respectively, these values were amended. Finally the categorical variables were converted to indicator variables to ready the data for machine learning.

Exploratory Data Analysis

Match success overview

Early analysis revealed that only 16.4% of dates resulted in a match. This was important to know as it meant we had imbalanced data. We would likely need to oversample the minority class (matches), undersample the majority class (non-matches), or generate synthetic samples when developing the machine learning model in order to combat the imbalance.

A unique identifier was created for each Person A within the dataset by identifying duplicates across the fields linked solely to Person A. This was then used to generate a histogram showing distribution of the number of matches attained by each individual.



The histogram shows that a significant proportion of people don't get any matches, reinforcing the problem we are facing. Potentially a machine learning model could help to pair up individuals on dates with other individuals that they are more likely to match with, which could reduce the proportion who leave the event without success.

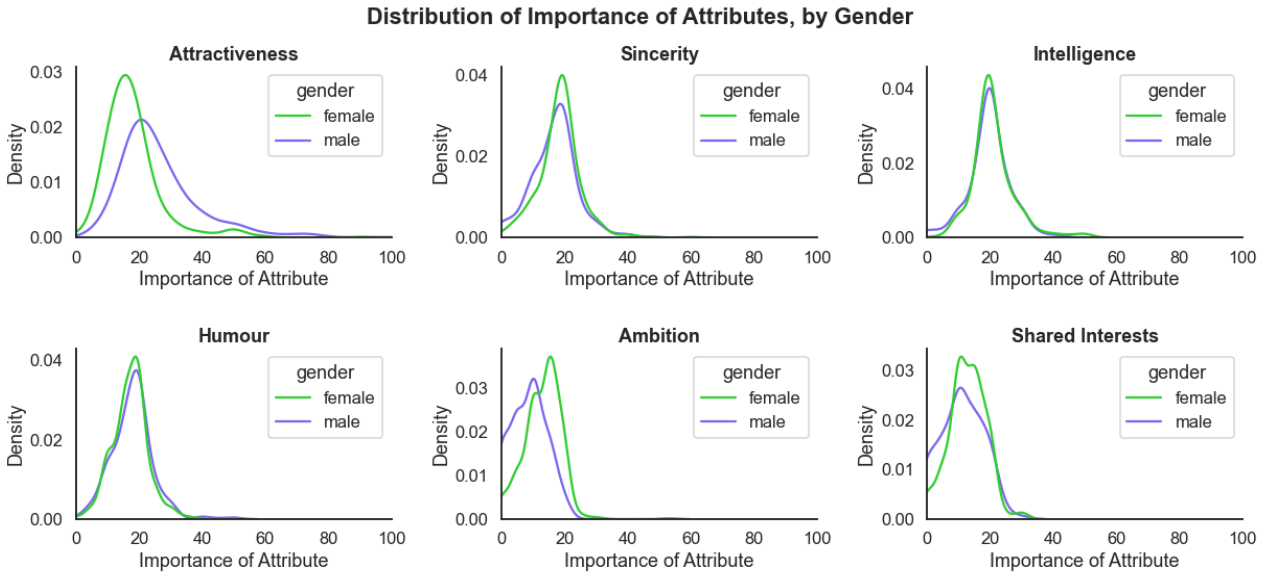
Age distribution

Below is a kde plot of age by gender.

This plot shows that most participants are in their twenties. On average the men are slightly older than the women, though not significantly. The age range is reasonably wide however, ranging from 18 to 55. If all participants go on dates with all other participants at the event, we could end up with some significant age differences, which could have a negative impact on the chance of a match. This is corroborated by a study in 2011 which showed that greater age disparity among speed dating participants tends to result in fewer matches (1). Therefore it could be worth creating some smaller 'sub-events' restricted to individuals within a particular age group.

Attribute analysis

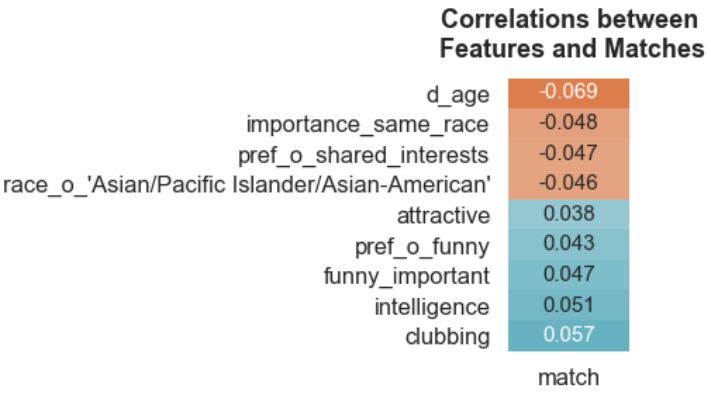
A similar plot was then generated for how importantly Person A rated six personal attributes. Note that each Person A was asked to assign an importance weighting to these six attributes such that the total was equal to 100%.



From the distribution plots above we can see that all attributes have a similar average weighting; there is no particular attribute that is significantly more or less valued that the others. This suggests a participant who is well-rounded across the attributes might have a better chance of a match than a participant who shows much of one attribute but little of the others. However some features have a greater spread of weightings - for example the attractiveness weightings range up to 100%, suggesting this attribute is more divisive of opinion, whereas the shared interests weightings only go up to 40%. We can also see that men and women tend to value certain attributes differently - for example, men tend to value attractiveness more, whereas women tend to be more interested in ambition.

Correlation between features and matches

Next is a heatmap of the strongest five positive and strongest five negative linear correlations between features in the dataset and the target variable, ‘match’.



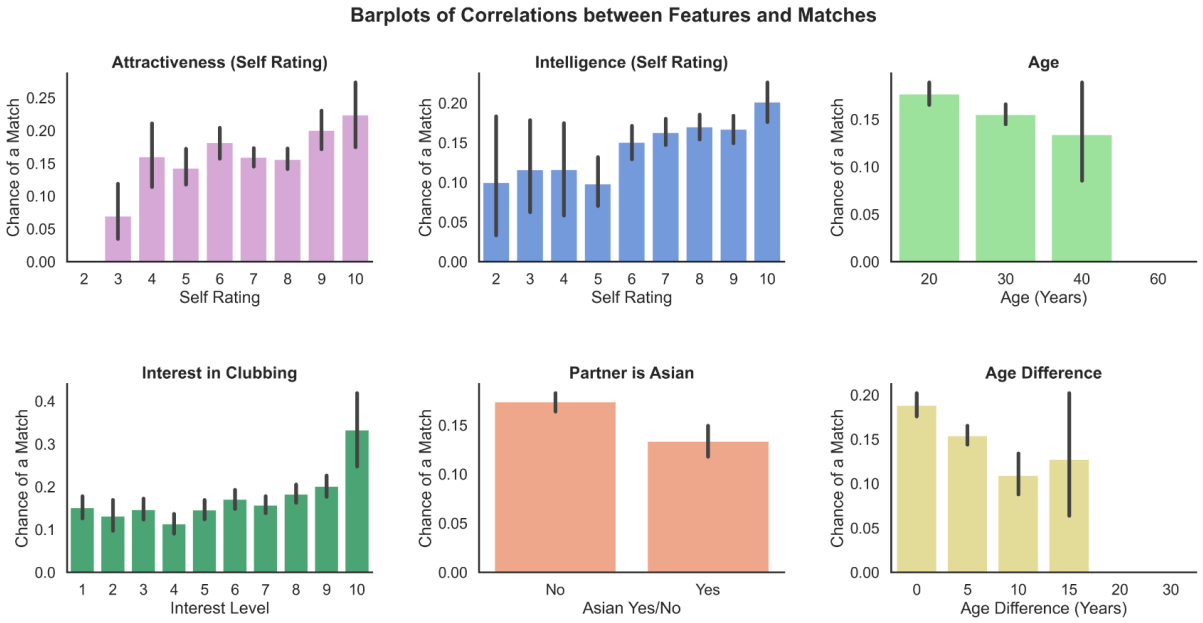
From this heatmap we can see that interest in clubbing has the strongest positive correlation with matches, suggesting that the more interested in clubbing you are, the more likely you are to leave the event with a match. This is not something that is traditionally linked to dating success, but perhaps it's because people who like to go clubbing a lot are typically quite outgoing and sociable, which should theoretically increase their chances on a date. Self-rating of intelligence shows a similar positive correlation, possibly as those who consider themselves to be intelligent may have more interesting opinions to share on a date, or a broader range of topics of conversation.

Conversely, the age difference between the two participants has the strongest negative correlation with matches, which makes sense as people of a similar age are more likely to have similar interests, experiences and lifestyles. How important someone values being the same race as their partner also has a strong negative correlation, which is

probably because they are more likely to write off a potential partner simply because they are a different race to them.

Note that all of these correlations are individually quite weak, so would be very poor predictors of a match on their own. This is where machine learning can come in as it can take into account multiple variables at once. It's also worth noting that the correlation coefficient used in the diagram only takes into account linear correlations, so there may still be non-linear correlations in some of the other variables that haven't been picked up. Again, supervised learning algorithms such as random forests are good at picking up on this kind of relationship.

The below plots show the average chance of a match across six different features, with 95% confidence intervals overlayed. The features plotted were chosen for their strong visual correlations.

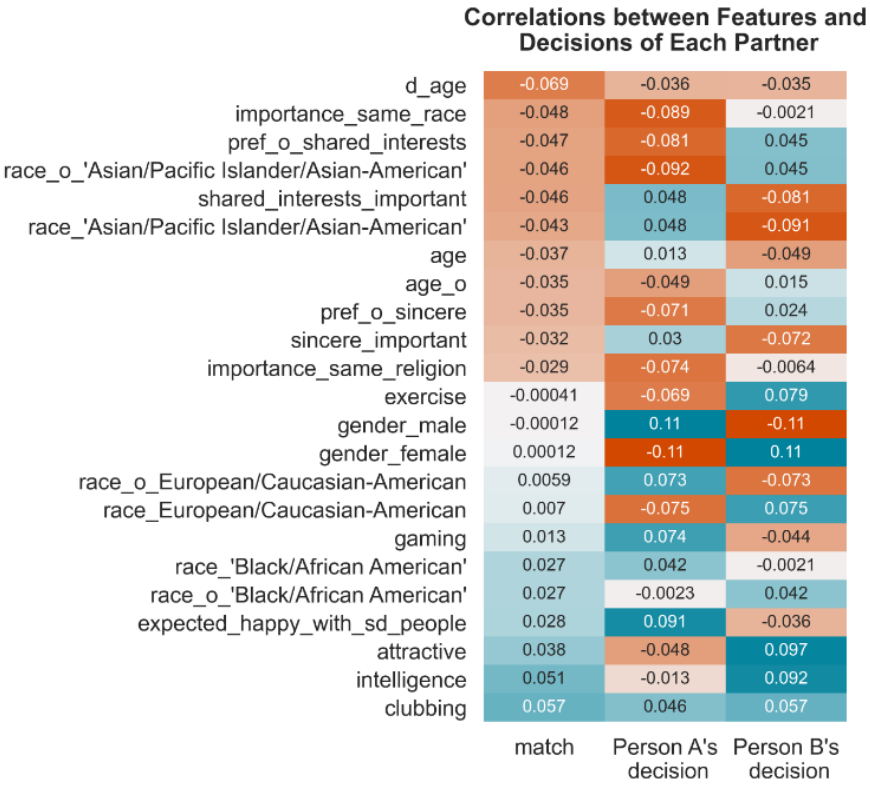


As expected from the heatmap of correlations, interest in clubbing, attractiveness and intelligence self-rating both show an upward trend, indicating a positive correlation with the chance of a match. Conversely age and age difference show a downward trend, reinforcing the negative correlations shown in the heatmap. For the clubbing and age difference plots in particular, we can see that some of the confidence intervals do not overlap, which shows that these features have a statistically significant impact on an individual's chance of a match. A two-sample t-test does indeed confirm that the chances of a match are different for those who rate their interest in clubbing as 10 versus everyone else, yielding a p-value of 2.49e-6.

Surprisingly, whether the individual's partner is Asian, Pacific Islander or Asian-American (referred to as Asian from now on) also has a statistically significant impact on the chance of a match, with Asian partners being only 13% likely to find a match compared to 17% for non-Asian partners. It's not immediately obvious why this might be, but it could be associated with cultural differences, or perhaps that people of Asian origin are more or less likely to be interested in certain hobbies that could affect their chance of a match. We also cannot rule out potential racial bias from some of the participants.

Correlation between features and individual decisions

Now let’s look at an extension of the heatmap of correlations we saw before; this time including not just correlations with matches, but also correlations with Person A's decision (i.e. whether or not they want to see their partner again), and correlations with Person B's decision in the rightmost column, in order to break down our analysis further.



We can see that certain features such as Person A's interest in clubbing appear to have a positive impact on both partner's decisions, and other features such as age difference have a negative impact on both decisions. However there are certain features that will raise one partner's likelihood of saying yes while lowering the other partner's likelihood of saying yes. The strongest such feature is gender, showing that male participants are more likely to say yes to their partner than female participants. This could be because women are more cautious in love because childbearing increases the negative cost of choosing the wrong partner (2). Another such feature is someone's self-rating of attractiveness; if they rate themselves highly, they are more likely to be approved of by their partner, but somewhat less likely to say yes back. Interest in exercise has a similar effect.

However the most common type of feature that appears to divide the partner's decisions appears to be race. For example, Asian participants are more likely to say yes to their partner but less likely to be said yes to. For white participants it is the opposite story - they appear to be more popular but also more selective. Black participants are typically more likely to say yes to their partner, but there is not much impact on their partner's decision.

The reasons behind these racial variances are unclear. One might hypothesise that people prefer to date those with the same race as them, which is why ethnic minorities have a lower chance of a match due to there being fewer of them. However 'samerace' is already a feature, and while it has a positive correlation with each partner's decision, the correlation is not as strong as that which is specific to the race of the participant. Therefore further research into these trends would be recommended in order to establish the reasons behind them, and if it is linked to racial prejudice, take steps to reduce this in future events.

Machine Learning

Preparation

To prepare the data for machine learning models, the data was scaled before being split into a training and test set. The random state was fixed for reproducibility; the data was shuffled given it was originally ordered by individual; and the labels were stratified in order to ensure the test and training data reflects the original dataset. Subsequently synthetic minority oversampling was applied to the training set to mitigate the imbalance between matches and non-matches.

Success criteria

When evaluating the performance of machine learning models, accuracy is a poor choice for this problem. This is because the data is imbalanced, so even if a model predicted "no match" for every couple, it would achieve an accuracy score of 84%. Instead we can use a confusion matrix to split the results by actual outcome (match / no match) and predicted outcome (match / no match). False positives are a problem for us because they mean participants would be predicted matches that turn out to be wrong, resulting in disappointment. However false negatives are also an issue because they would result in missing potential matches between participants. Therefore it makes sense to use a metric such as the F1 score, which takes into account both false positives and false negatives by computing a weighted average of precision and recall.

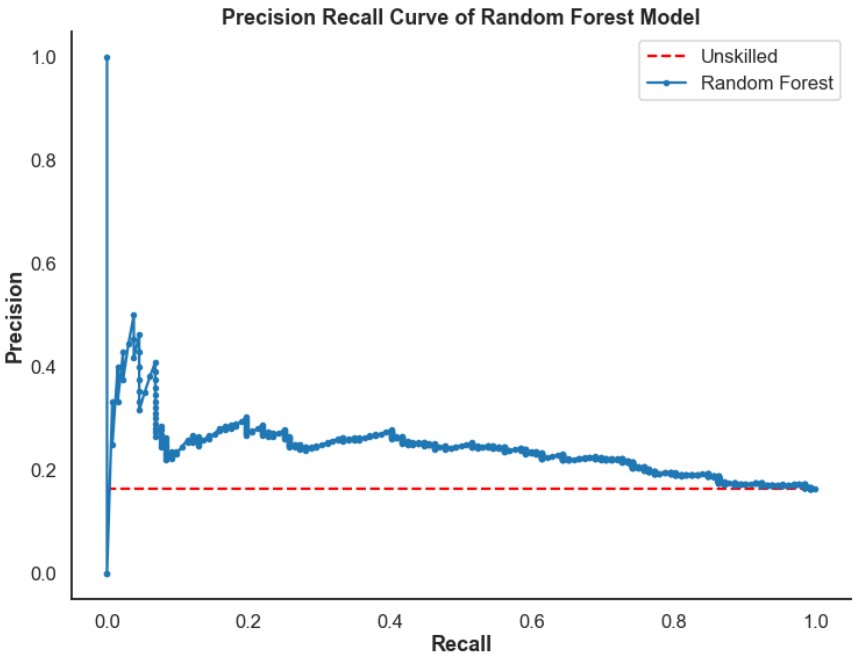
Model selection

A number of different models were tried out to start with and evaluated using cross validation on the F1 score. The random forest classifier came out on top. This is unsurprising for this type of problem, because tree-based models are able to capture nonlinear relationships between features and labels, as well as interaction effects. Hyperparameter tuning identified an optimal random forest classifier with 70 estimators and each leaf containing at least 1% of the training data. This model was then used to predict the test set data and evaluated using the F1 score.

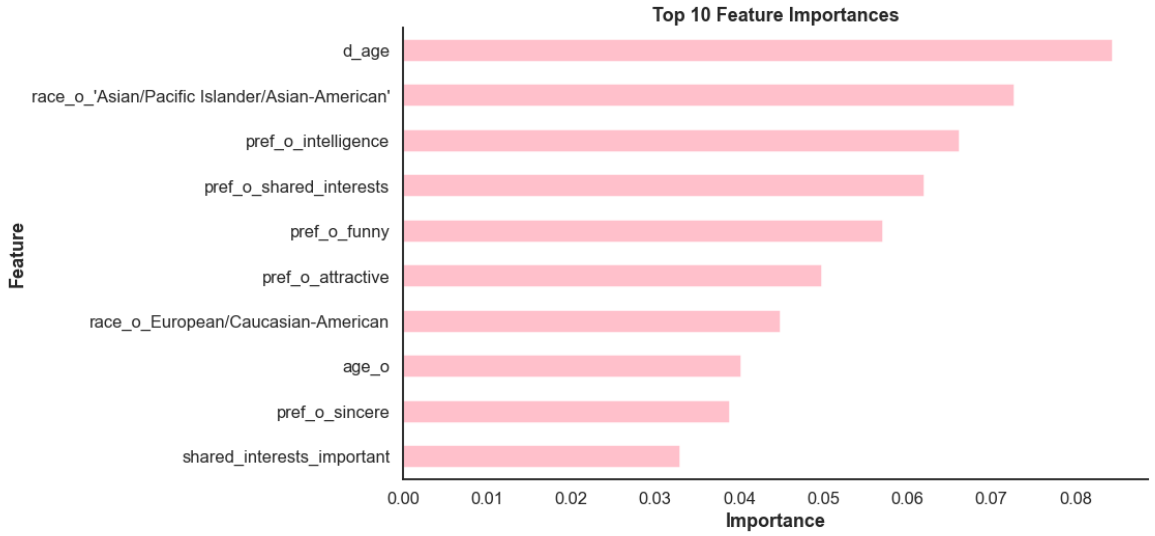
Model evaluation

Even though a suitable model and optimal parameters had been used, the resulting predictor produced poor results, with an F1 score of only 0.25. This is likely a combination of romantic matches being notoriously hard to predict, and also the limited nature of the data we are using. In particular we have very little information about Person B. In order to improve the model I would recommend collecting further data - Person B should be asked the same questions as Person A, and in addition a number of extra questions could be included, such as physical attributes/preferences, education levels, values, political leanings and additional interests such as animals or travelling.

The precision recall curve below further demonstrates that the model is relatively weak, as the plot of the random forest classifier does not get near to (1,1). However it does show that it clearly outperforms an unskilled model, shown by the dashed red line.



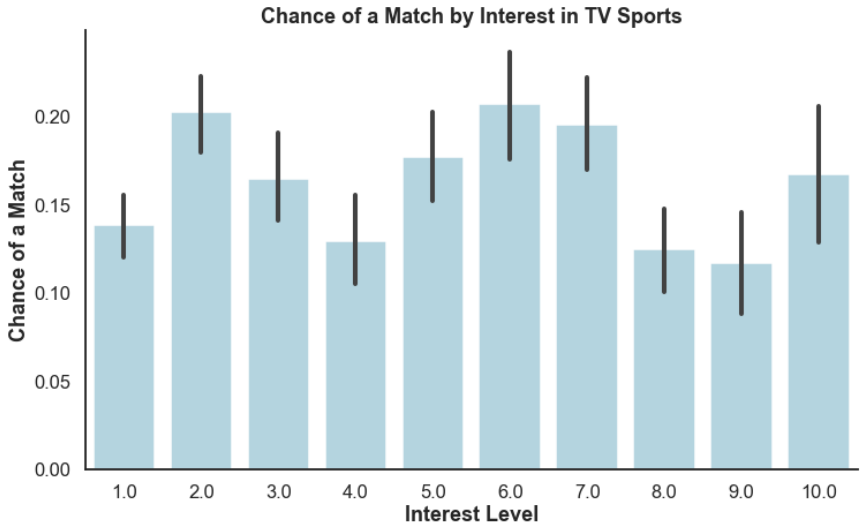
Below is a plot of the 10 most important features in the random forest classifier.



As expected, the features that prove most important in predicting matches roughly correspond to those features with a strong linear correlation to matches, which we identified in the Exploratory Data Analysis section. For example we can see age difference is the most important feature, which is in line with our heatmap showing it to have the strongest correlation to matches of all the variables. However some other variables with strong linear correlations are missing. This could be because there are relationships between the features themselves; for example, if two features are highly correlated the model could just use one of them for its predictions, effectively discarding the other feature as it doesn't add any extra value.

Another point of interest is that most of the top 10 important features relate to Person B. This is possibly because there is very little data on Person B compared to Person A, and so any data that the model can use on Person B will be weighted much more strongly.

Finally there are some variables on the feature importances diagram that didn't show any linear correlation in our earlier analysis, such as interest in TV sports. This is likely because there is a relationship between interest in TV sports and chance of a match, but it is nonlinear, so didn't get identified by the Pearson correlation coefficient. As mentioned previously, the ability to identify nonlinear relationships between features and values is one of the advantages of tree-based models. A plot of match likelihood by interest in TV sports is shown below to demonstrate this. Note that there is no obvious linear trend, but many of the confidence intervals do not overlap, suggesting that the difference in match likelihood by interest in sport is statistically significant.



Conclusions and Recommendations

Conclusions

The aim of this project was to investigate which factors affect the likelihood of a match, and to use supervised machine learning techniques to develop a predictive model that aims to predict whether or not two participants will be a match.

Cleaning the data highlighted several data quality issues. A number of missing values were present in the data, which had to be removed. There were some anomalous values such as 13 when the answer should have been on a scale from 0-10. There were also a number of spelling mistakes in the feature names which made them difficult to refer to in the code, as well as a few feature names that didn't have a description. Additionally many of the features were redundant for the purposes of this analysis, typically because they weren't possible to collect in advance of the date.

Exploratory data analysis helped to identify key features that affect a couple's chance of matching, by calculating and visualising the linear correlation between the features and the likelihood of a match. We found that certain features such as interest in clubbing and intelligence self-rating had a positive impact on the chance of a match, whereas features such as age difference had a negative impact. In some cases the difference in average chance of a match across responses to a particular question was shown to be statistically significant. Further analysis highlighted the complexity of the problem, by demonstrating that some variables would increase the chance of Person A saying yes to their partner, but would decrease the chance of Person B saying yes. The most common such variables are related to race.

To predict the matches, a random forest model was fit to the data and optimised using hyperparameter tuning. The F1 score was used to evaluate the model, given it works well on imbalanced data and takes into account both false positives as well as false negatives. A plot of the precision recall curve of the model showed that it outperformed an unskilled model, and therefore is a better predictor of matches than the current method where participants are simply paired up with everyone else at the event. Therefore if we incorporate this predictive model into our events going forwards, we could selectively pair up individuals based on their likely matches, increasing match likelihood and customer satisfaction. However it’s worth noting that the resulting model achieved an F1 score of 0.25 which is fairly weak. This is likely because there is very little information about each participant in the dataset; especially Person B.

Recommendations

Steps should be taken to make the dataset cleaner for future analysis. When collecting the data electronically, automatically preventing the questionnaire from submitting until all answers are filled in would be advised to prevent missing values. The values that can be entered in response to each question should be restricted, for example via a drop-down list, to eliminate outliers. Spelling mistakes in the feature names should be fixed to aid future analysis. Questions that don’t add any value (e.g. those that can only be asked after the date) could also be dropped for this specific prediction analysis.

To improve the predictive power of the machine learning model, increasing the data collected would be advised. More data should be collected in advance of the date on Person B to match the information collected on Person A, and additionally extra data could be collected, such as physical attributes/preferences, education levels, values, political leanings and additional interests. This is likely to give a more complete picture of each individual and result in a stronger predictive model.

To make use of the predictions made by the random forest classifier, data could be collected from a large number of individuals, matches could be predicted using the model, and then smaller speed dating events could be organised where the participants are grouped according to their likely matches. This would ensure that participants get to see their most likely matches and spend less time talking to people that they don’t feel a connection with. In turn this would lead to more participants leaving with a match, boosting the business’s reputation and thereby attracting new customers and revenue.

References

(1) "Science of Speed Dating Helps Singles Find Love" *Scientific American*, 2012

(2) "Speed dating: Why are women more choosy?" *BBC News*, 2014