Machine Learning Model Summary and Interpretation

Objective

Our goal was to build a classification model to **predict whether two individuals would not match** based on a set of behavioral and personal attributes. The dataset comprised **9000 samples** and **123 features**. Given class imbalance and the nature of the problem, our primary evaluation metric was **recall**, especially for the minority class (non-matches).

Introduction

Everyone has gone on a bad date, and it's hard not to take it personally. Humans have taken rejection so personally that entire research fields have been dedicated to understanding why people connect—or why they don't. I was curious: using machine learning, could I predict whether two people would not go on a second date? This report explores that question by building predictive models trained on real-world speed dating data.

This could be useful for anyone who has ever been curious about dating compatibility—whether you're navigating the dating world yourself or trying to design technology to help others find love (or avoid mismatches).

In brief: we used statistical analysis and machine learning to identify the key factors that determine whether two people are unlikely to match.

Dataset

We used the Speed Dating dataset from Kaggle, which contains information collected during real-world speed dating events. Participants rated each other on attributes like attractiveness, intelligence, humor, and shared interests.

The final dataset included 8,293 samples and 15 engineered and cleaned features. We focused on features that were both statistically significant and interpretable. Notably, 84% of the outcomes were non-matches, and only

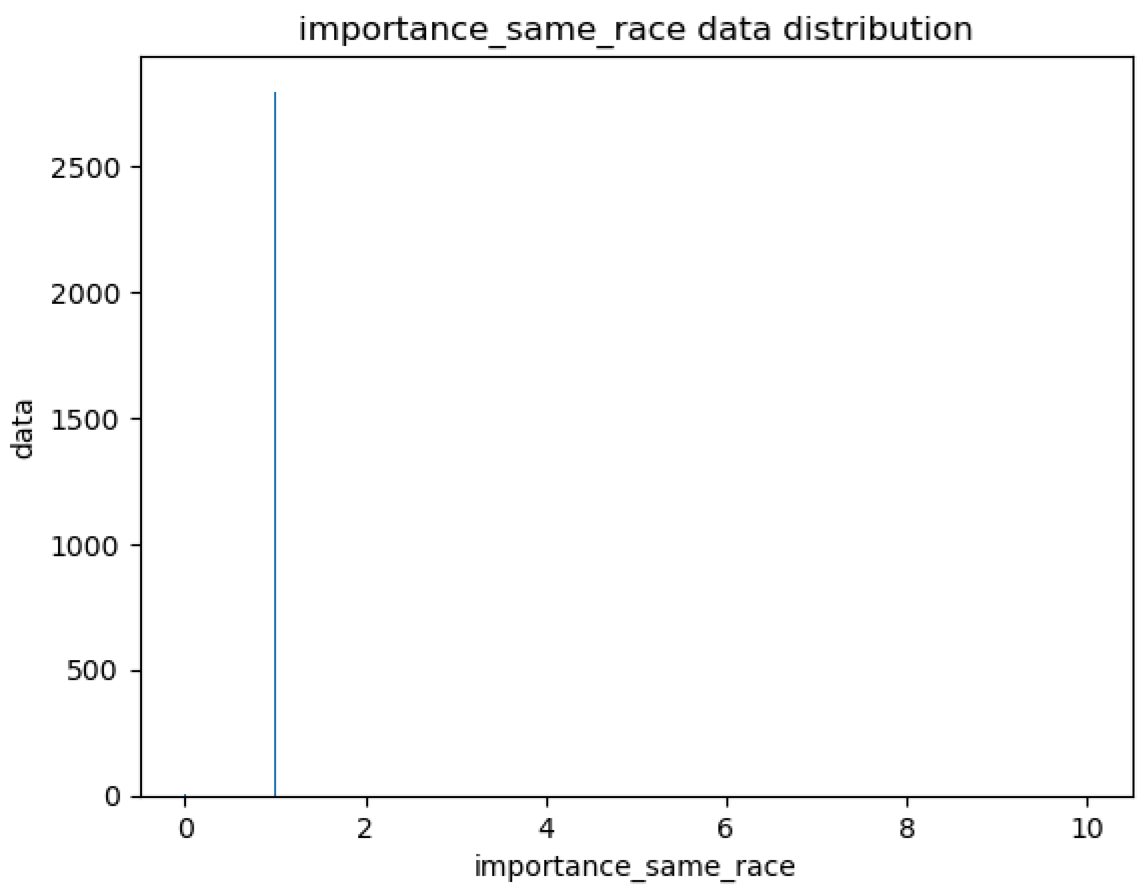
16% were matches—creating a class imbalance that had to be accounted for in model design.

Key preprocessing decisions included:

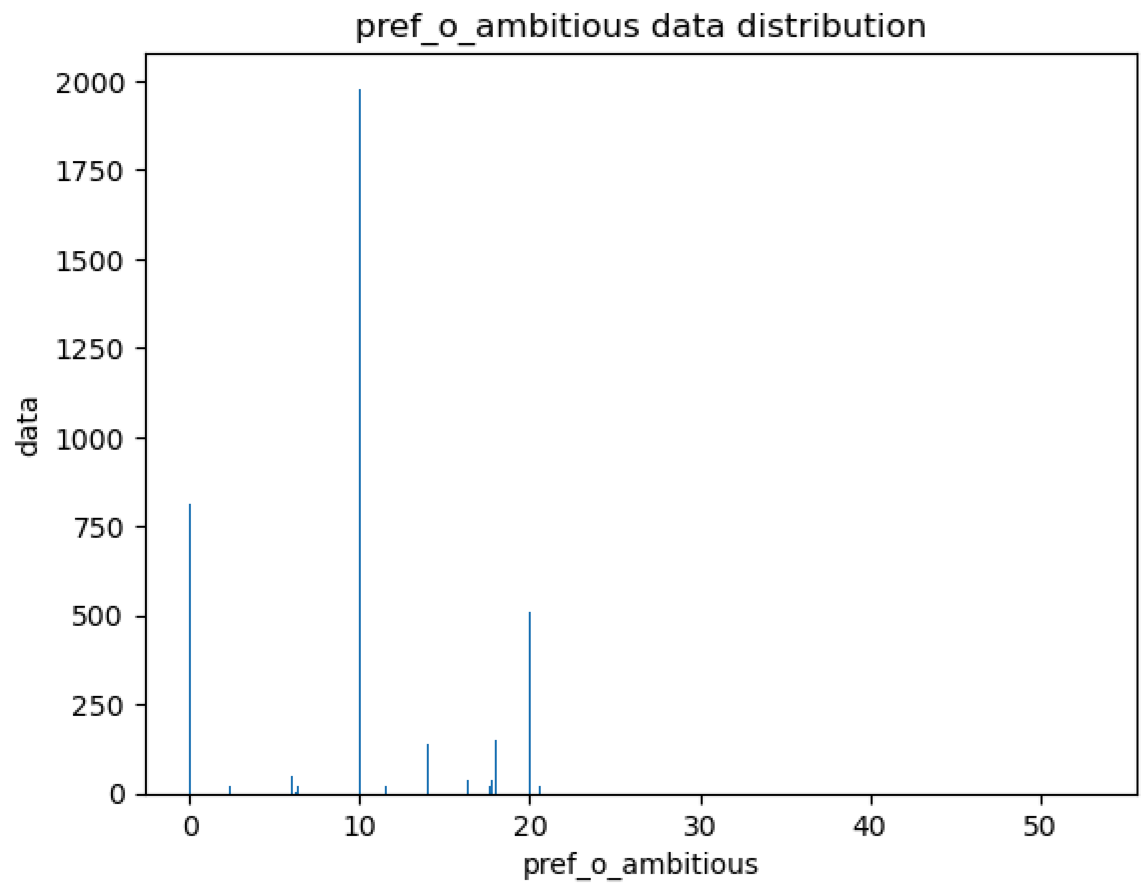
- **Removing improperly formatted or one-dimensional columns (e.g., entries like `[2-5]`)**
- **Using effect size (Cohen's d) and p-values to retain only the most meaningful features**
- **Applying class balancing techniques and threshold tuning to improve recall for both classes**

Exploratory Data Analysis (EDA)

Univariate Charts To understand the data distributions, we created histograms for key variables.



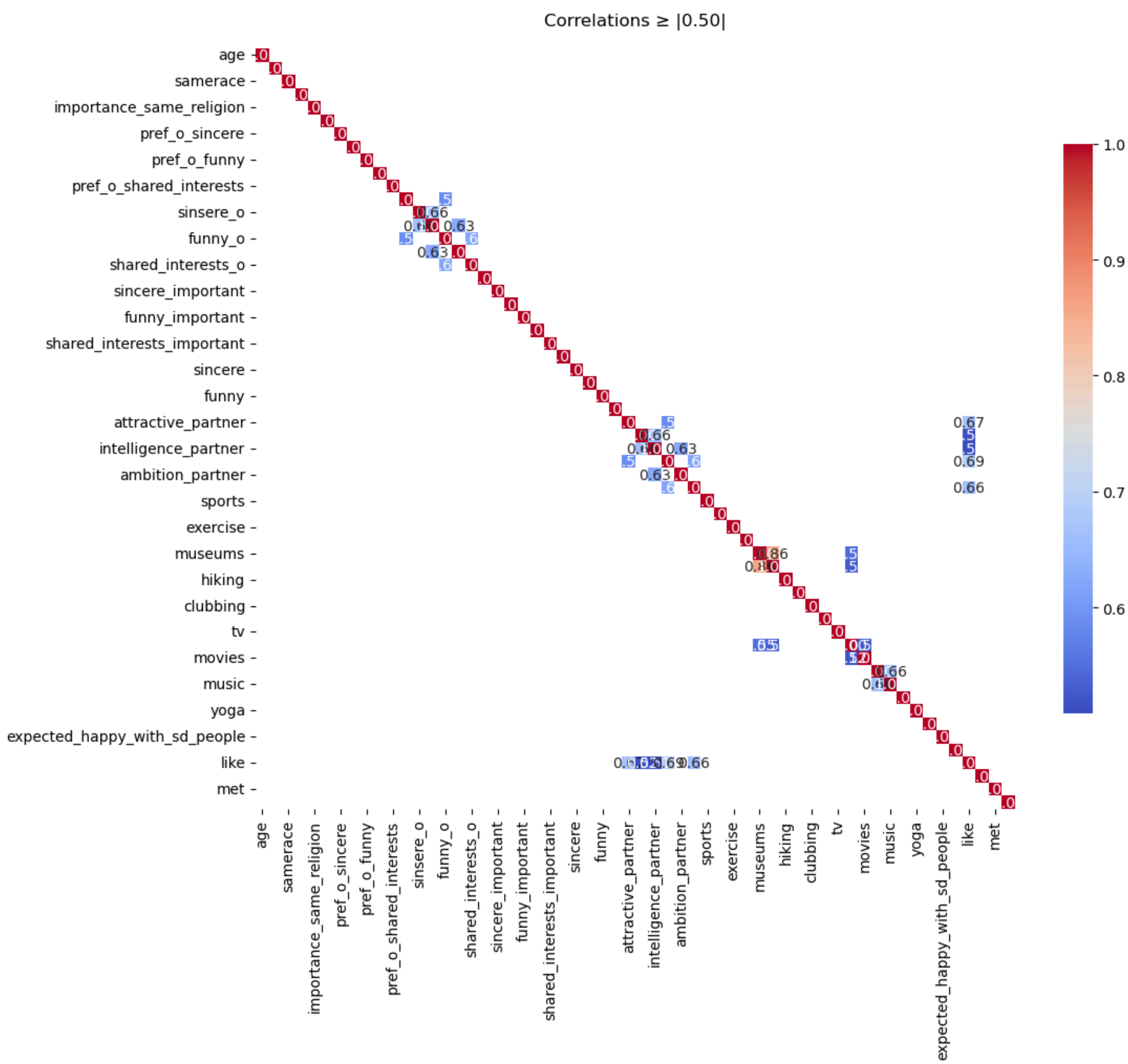
- **The histogram of `importance_same_race` revealed a sharp spike around a single value, implying a lack of variability. This was unexpected and signaled the need to remove it due to low informational value.**



- **Conversely, the distribution of `pref_o_ambitious` scores showed a wider range of responses between 0 and 20, confirming it as a more meaningful and potentially predictive feature. This also raised issues of being above the iqr range and having outliers so we had to take away values of the data set acordingly**

Takeaway: Features with highly skewed or constant distributions were reviewed and considered for removal if they lacked interpretability or variance.

Correlation Heatmap



A correlation matrix was generated to investigate multicollinearity between features. To avoid over-cluttering, we visualized only correlations above |0.50|.

- **Strong correlations were found between partner and self-ratings like `attractive_partner` and `attractive_o`, and `shared_interests_o` and `shared_interests_partner`, which was expected.**
- **These clusters indicated potential multicollinearity but also reaffirmed that these paired features reflect reciprocal impressions — valuable in a dating context.**

Takeaway: The filtering approach allowed clear insight into redundant features while avoiding noise.

Data Cleaning from Distribution Charts

Some columns, such as those with values in the `[2-5]` format, had inconsistent or broken formatting. These caused failed plots and empty histograms, leading to the discovery that such columns were either improperly parsed or categorically uninformative.

- **As seen in the data distribution charts (e.g., for `importance_same_race`), several variables appeared completely one-dimensional or visually blank.**
- **These were removed from the dataset after confirming their irrelevance during visualization.**

Takeaway: EDA helped identify structural formatting issues in the dataset that were invisible from basic `.head()` inspection. These columns were dropped.

Bivariate Analysis & Statistical Significance

We assessed feature relevance using Cohen's d effect size and p-values from hypothesis testing. Statistically significant features ($p < 0.05$) were retained for modeling. This approach gave us confidence in the features' influence on the target variable.

Below are the selected features, all with p-values = 0.000 and their corresponding Cohen's d effect sizes:

Feature	p-value	Cohen's d
attractive_o	0.000	0.720
sinsere_o	0.000	0.444
intelligence_o	0.000	0.457
funny_o	0.000	0.766
ambitous_o	0.000	0.374
shared_interests_o	0.000	0.741
attractive_partner	0.000	0.715
sincere_partner	0.000	0.443
intelligence_partner	0.000	0.456
funny_partner	0.000	0.767
ambition_partner	0.000	0.373

shared_interests_partner	**0.000**	**0.736**
expected_num_matches	**0.000**	**0.360**
like	**0.000**	**0.852**
guess_prob_liked	**0.000**	**0.703**

> **Takeaway: These features demonstrated both statistical significance and large effect sizes, justifying their inclusion in the final model. Notably, features like `like`, `funny_partner`, and `shared_interests_o` not only aligned with psychological literature but also had high predictive power. Ultimately we had a final data shape of** `(8293, 15)`

Feature Importance

Feature importance was evaluated using the best-performing Random Forest model. Below are the most influential features:

Feature	Importance
like	0.1141
funny_o	0.1032
shared_interests_o	0.1032
attractive_o	0.1028

attractive_partner	0.0968
funny_partner	0.0858
expected_num_matches	0.0799
shared_interests_partner	0.0746
guess_prob_liked	0.0722
ambitious_o	0.0606
sincere_partner	0.0547
ambition_partner	0.0521

Interpretation:

- Features like **'like'**, **'funny_o'**, and **'shared_interests_o'** ranked highest, which aligns with intuition: a participant's perception of the other person's humor and shared interests are central to connection.
- **Humor** was statistically insignificant in logistic regression but was kept due to its theoretical relevance from literature.

Feature Selection Process

A combination of statistical analysis (logit model coefficients and p-values) and theoretical judgment (literature on relationship psychology) was used to select features. Features with p-values > 0.05 were considered for removal, but some were retained if their removal did not yield meaningful performance gain.

Modeling Overview

Model Name	Best Parameters	Optimal Feature Set	ROC AUC (Test Set)
Logistic Regression	{'C': 1, 'penalty': 'l2', 'solver': 'lbfgs'}	Scaled / Full (15)	0.821

Random Forest	{'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}	Unscaled / Full (15)	0.825
SVM	{'C': 1, 'kernel': 'rbf', 'gamma': 'scale'}	Scaled / Full (15)	0.814

Threshold tuning was applied post-modeling to optimize recall.

Classification Reports *(all metrics in %)*

Logistic Regression (Tuned Threshold):

Class	Precision	Recall	F1-Score
0	91.0	89.0	90.0
1	51.9	56.4	54.0

Macro Avg Recall: 73.0%
Weighted Avg Accuracy: 84.0%

Random Forest (Tuned Threshold):

Class	Precision	Recall	F1-Score
0	91.0	87.0	89.0
1	48.4	59.9	53.5

Macro Avg Recall: 73.0%
Weighted Avg Accuracy: 82.0%

SVM (Tuned Threshold):

Class	Precision	Recall	F1-Score
0	92.0	86.0	89.0
1	47.5	60.5	53.2

Macro Avg Recall: 73.0%
Weighted Avg Accuracy: 82.0%

Conclusion

Major Findings

- All three tuned models converged on a **macro recall of 73%**, suggesting a well-balanced detection capability across classes.
- **Random Forest** had the highest ROC AUC (0.825) and robust performance across metrics.
- Feature importance aligned with expectations from both domain knowledge and model-driven analysis.

Use Case Alignment

This model supports platforms or studies that aim to **identify non-match pairs** early, allowing for optimized pairing algorithms or interventions.

Next Steps

- Investigate ensemble or boosting methods with stricter recall focus.
- Expand dataset to improve generalizability and handle rare behavioral patterns.
- Include more nuanced behavioral/psychological indicators beyond structured numeric inputs.

Assumptions/Future Work

- Thresholds were tuned for recall maximization — use-case-specific rebalancing may be required.
- Time-sequence modeling and interactive feedback loops (e.g., reinforcement learning) may improve performance.

- Additional research could compare results across demographic splits or with real-world dating app outcomes.