Thanks, Cupid

Fundamentals of Data Science (CSC4780)
Fall 2022 Dr. Berkay Aydin
Georgia State University
12/06/2022

Cupid Scientists: Cindy Thai, Lorena Burrell, Capella Edwards, Venkata Mani Mohana Rishitha Srikakulapu, Laurel Sparks

Overview

Introduction

- Business Understanding
- Data Understanding (Sources, Exploration, & Preprocessing)

Model Selection

- Model Selection
- Feature Selection
- Data Sampling
- Model Optimization

Evaluation

- Performance Metrics

Conclusion

- Summary
- Recommendation

Introduction (Business Understanding)

Dating Applications:
Generate a list of suitable partners for the person looking for love to connect with based on location and limited filtering

Goal: Sell premium services to allow the user to generate a more refined list of matches that have been filtered based on the users preferences.

Businesses will benefit from:

- predicting which people are most likely to match with each other
- Predicting what matches have a high probability that the users will go on a date.

This will yield **more profits** for the business to **sell more premium services**, and get more users to join the app which will **increase the variety of dating partners** for paying customers.

Data Sources and Data Exploration (Data Understanding)

- Data sourced from a Speed Dating experiment conducted by Columbia Business School (Raymond Fisman, Sheena S. Iyengar, Emir Kamenica, and Itamar Simonson, 2006)
  - Wide range of descriptive features (continuous, categorical) & 8,000+ instances
- After Data Exploration we selected **19** out of 195 features to base our model on

Feature	Desc.	Count	% of Missing	Card.	Min.	Q1	Median	Q3	Max.	Mean	Std. Dev.
hobby_diff_phys	sum of difference between hobby/interest value...	8188	0.02	36	0.00	8.00	12.00	15.00	35.0	12.01	4.89
hobby_diff_out	sum of difference between hobby/interest value...	8188	0.02	52	3.00	12.00	16.00	21.00	58.0	17.23	6.77
hobby_diff_in	sum of difference between hobby/interest value...	8188	0.02	38	2.00	12.00	16.00	19.00	41.0	15.95	5.27
attr_diff	difference in self-rated amount vs partner's p...	8136	0.03	758	0.67	19.00	26.50	37.00	131.0	30.56	16.15
sinc_diff	difference in self-rated amount vs partner's p...	8136	0.03	726	1.00	15.00	19.48	24.00	62.0	20.00	8.02
intel_diff	difference in self-rated amount vs partner's p...	8136	0.03	633	0.00	18.98	23.00	29.00	69.0	24.39	8.93
amb_diff	difference in self-rated amount vs partner's p...	8100	0.03	702	0.00	7.50	11.00	15.00	57.0	11.55	5.57
fun_diff	difference in self-rated amount vs partner's p...	8118	0.03	640	0.00	15.50	20.00	24.00	61.5	20.24	7.65
income_diff	difference between incomes	2178	0.74	1061	8.00	6591.00	14997.00	26150.00	85670.0	18447.40	15078.11
age_diff	difference between ages	8159	0.02	25	0.00	1.00	3.00	5.00	32.0	3.66	3.06
confidence	percentage of people each person dating expect...	1790	0.79	50	0.00	0.15	0.25	0.38	20.0	0.39	0.93
exphappy	user expectation of happiness with speed datin...	8245	0.01	11	1.00	5.00	6.00	7.00	10.0	5.52	1.72
out_freq	rating of how often user goes out (not necessa...	8267	0.01	8	1.00	1.00	2.00	3.00	7.0	2.16	1.11
date_freq	rating of how often user goes on dates	8249	0.01	8	1.00	4.00	5.00	6.00	7.0	5.02	1.44
imprace	importance of having same racial/ethnic backgr...	8267	0.01	21	0.50	2.00	3.50	5.00	10.0	3.79	2.04

Feature	Desc.	Count	% of Missing	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
samerace	are the two participants the same race	8346	0.0	2	0	5039	60.38	1	3307	39.62
same_goal	whether both people have the same goal in part...	8346	0.0	2	0	5805	69.55	1	2541	30.45
same_career	whether both intended career paths fall into t...	8346	0.0	2	0	6852	82.10	1	1494	17.90
match	target: did they end up matching	8346	0.0	2	0	6972	83.54	1	1374	16.46

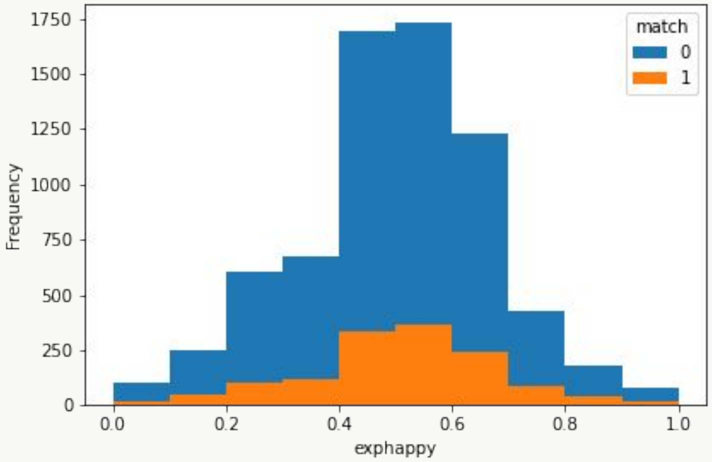
Figure 1 & 2: Data Quality Reports for Continuous & Categorical features (used for data exploration and feature selection)

Data Preprocessing

- **Handling Missing Values:**
  - Continuous features: mean imputation
    - Income feature (‘income_diff’): KNN imputation (better estimate, far more missing values)
  - No missing values in categorical features
- **Handling Outliers:**
  - Dropping instances with outliers would remove near 50% of all instances
  - Clamped outliers with Tukey’s Range Test & Interquartile-Range
  - Clamped ‘income_diff’ with 0.05 & 0.95 percentiles (due to high amount/severity of outliers)
- **Normalization:**
  - Normalized all continuous features into [0,1] range normalization

Data Transformation

- Reduced 195 features to 19 features (18 descriptive; 1 target)
- Handled missing values with mean imputation or KNN imputation
- Normalized continuous features with range normalization
  - Preserves original relationships of original feature distributions while putting features into normalized range for model building



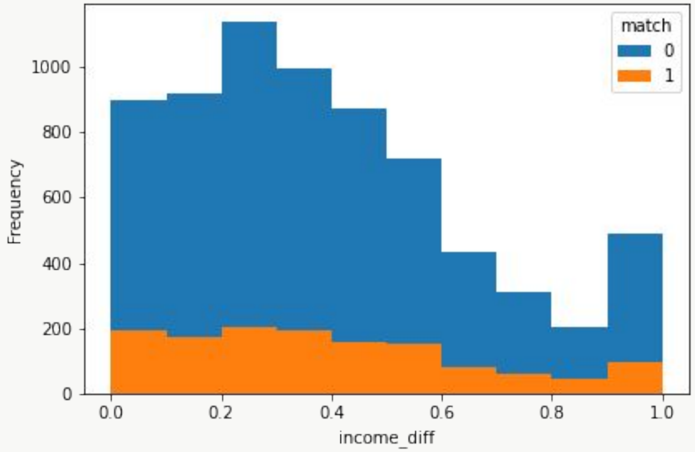


Figure 3 & 4: Example histogram distributions of normalized features

Model Selection

Decision Tree

Information-Based
Highly interpretable (gives businesses choice in course of action based on decision rule)

kNN

Similarity-Based
Lazy Learner (memorizes training data, meaning it takes no time in the training phase)

Naive Bayesian

Probability-Based
Fast implementation (no iterations needed)
Highly scalable

Feature Selection

```
kNN accuracy score: 0.81792
```

```
kNN accuracy score: 0.83421
```

Impurity-Based Univariate Feature Selection (IUFS)

- Utilized Entropy, Gini Index, Information Gain Ratio.

Recursive Feature Elimination (RFE)

- Using a logistic regression model to rank descriptive features

Based on results of IUFS and RFE: reduced 18 descriptive features to **16 features**

- Removed ‘same_goal’ (same goal in speed dating; removed due to IUFS), ‘amb_diff’ (correlation between self-rating of ambition; removed due to RFE)

Tested **all** features and **selected** features on kNN classifier: accuracy score **improved** with selected features

Data Sampling

The remaining 25% of the data was used for testing

75% of the data was partitioned for training

Model Optimization

Each model had its parameters optimized via an accuracy-based grid search. Then we found the optimal threshold based on the F1 Score, Gilbert Skill Score (GSS), and Hanssen-Kuipers Skill Score (TSS) evaluation of each model. All thresholds were low (biased towards positive predictions).

Gaussian Naive Bayes:

- Performed best with smoothing variance of 0.01 (0.01 of largest variance of all features added to all variance calculations)

Decision Tree:

- Performed best with maximum depth of 10, purity criteria using gini index over entropy or log loss

K Nearest Neighbors:

- Performed best with distance-based weights, Manhattan distance, and leaf size of 10

	Actual Positive	Actual Negative
Predicted Positive	166	578
Predicted Negative	193	1150

	Actual Positive	Actual Negative
Predicted Positive	164	241
Predicted Negative	195	1487

	Actual Positive	Actual Negative
Predicted Positive	232	194
Predicted Negative	127	1534

Figure 5: Confusion matrices for selected models

Model Evaluation

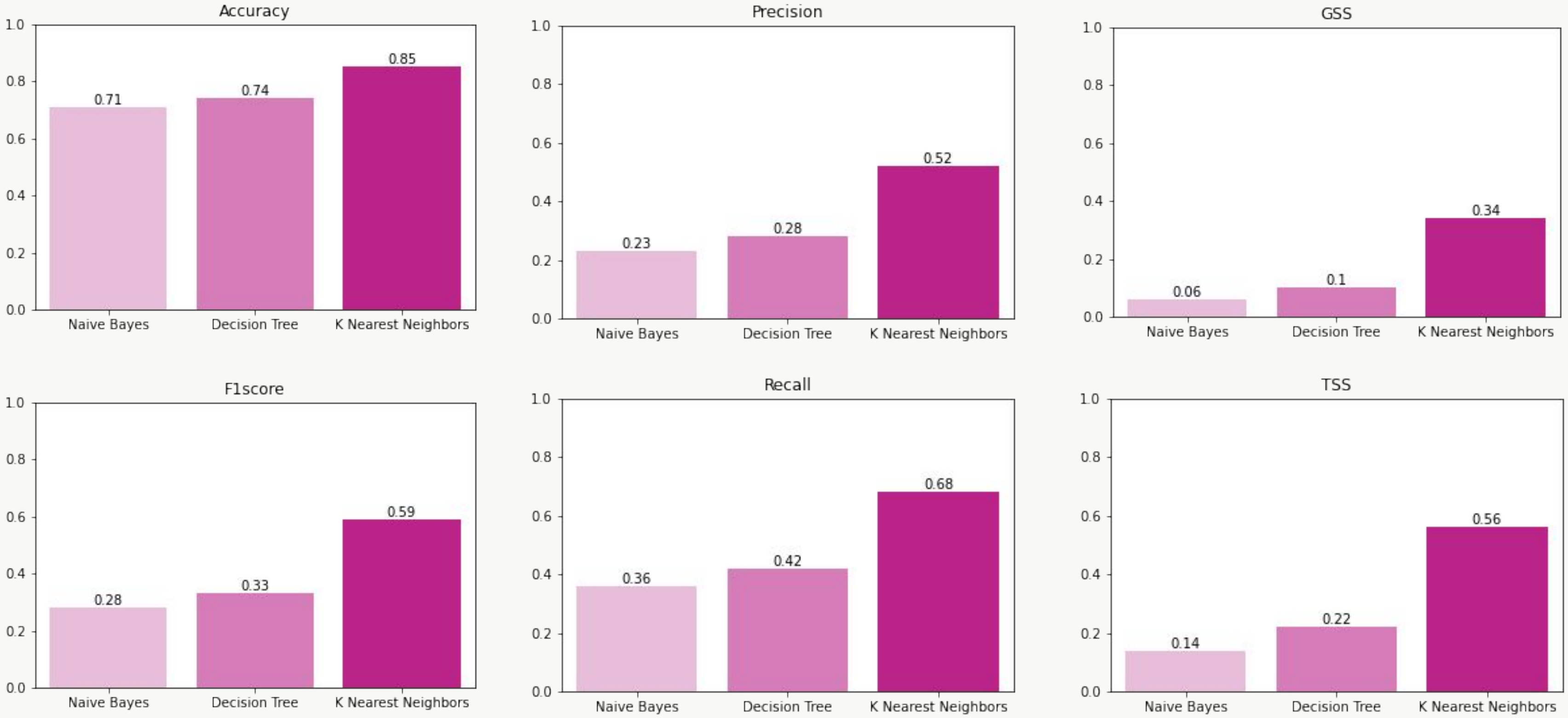


Figure 6: Selected models and performance measures

Conclusion

We recommend that our dating application used the **K Nearest Neighbors (kNN)** model with our **optimized hyperparameters** for predicting matches.

- kNN had the highest accuracy score (0.85) and highest score for every performance metric tested
- kNN model has low optimal threshold: biased toward positive predictions (matches)
  - More positive matches = more user engagement
- In scenarios with multiple matches; the most ideal match will have the highest probability score
- **We can sell these services behind a premium subscription and generate profit**