Supplemental document

Prediction of frozen gain episodes using natural language processing

Seibi Kobara

Table S1. Model performance results comparing different combinations of machine learning algorithms and feature sets.

Classifier	Hyperparameters	Feature set	F1 score of micro average	F1 score of macro average	Accuracy
Decision tree	{}	demographics	0.82828283	0.82828283	0.82828283
XGBoost	{'classifier__max_depth': 6, 'classifier__n_estimators': 10}	demographics	0.82828283	0.82828283	0.82828283
XGBoost	{'classifier__max_depth': 6, 'classifier__n_estimators': 50}	demographics	0.82828283	0.82828283	0.82828283
XGBoost	{'classifier__max_depth': 10, 'classifier__n_estimators': 10}	demographics	0.82828283	0.82828283	0.82828283
XGBoost	{'classifier__max_depth': 10, 'classifier__n_estimators': 50}	demographics	0.82828283	0.82828283	0.82828283
KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 10}	demographics	0.82828283	0.82828283	0.82828283
SVM	{'classifier__C': 5, 'classifier__kernel': 'sigmoid'}	ngram	0.82828283	0.82241216	0.82828283
Logistic regression	{}	cluster	0.82828283	0.82651273	0.82828283
XGBoost	{'classifier__max_depth': 10, 'classifier__n_estimators': 50}	ngram	0.81818182	0.81590909	0.81818182
SVM	{'classifier__C': 5, 'classifier__kernel': 'rbf'}	cluster	0.81818182	0.81816327	0.81818182
SVM	{'classifier__C': 10, 'classifier__kernel': 'rbf'}	cluster	0.81818182	0.81816327	0.81818182
SVM	{'classifier__C': 5, 'classifier__kernel': 'sigmoid'}	ngram	0.80612245	0.79855025	0.80612245
XGBoost	{'classifier__max_depth': 6, 'classifier__n_estimators': 50}	ngram	0.80612245	0.80364863	0.80612245
Logistic regression	{}	ngram	0.7979798	0.7842197	0.7979798
XGBoost	{'classifier__max_depth': 10, 'classifier__n_estimators': 10}	cluster	0.79591837	0.79377104	0.79591837
SVM	{'classifier__C': 1, 'classifier__kernel': 'linear'}	word2vec	0.79591837	0.7951505	0.79591837
SVM	{'classifier__C': 5, 'classifier__kernel': 'linear'}	word2vec	0.79591837	0.7951505	0.79591837
Naive Bayes	{}	ngram	0.78787879	0.78779218	0.78787879
Logistic regression	{}	word2vec	0.7755102	0.77466555	0.7755102
XGBoost	{'classifier__max_depth': 10, 'classifier__n_estimators': 10}	ngram	0.75757576	0.75735294	0.75757576
Naive Bayes	{}	ngram	0.75757576	0.75332226	0.75757576
XGBoost	{'classifier__max_depth': 10, 'classifier__n_estimators': 50}	cluster	0.75757576	0.75635767	0.75757576

SVM	{'classifier__C': 1, 'classifier__kernel': 'linear'}	demographics	0.75757576	0.75635767	0.75757576
SVM	{'classifier__C': 5, 'classifier__kernel': 'linear'}	demographics	0.75757576	0.75635767	0.75757576
SVM	{'classifier__C': 10, 'classifier__kernel': 'linear'}	demographics	0.75757576	0.75635767	0.75757576
Logistic regression	{}	demographics	0.75757576	0.75635767	0.75757576
XGBoost	{'classifier__max_depth': 6, 'classifier__n_estimators': 10}	cluster	0.75757576	0.75555556	0.75757576
SVM	{'classifier__C': 5, 'classifier__kernel': 'sigmoid'}	ngram	0.74747475	0.74070194	0.74747475
SVM	{'classifier__C': 10, 'classifier__kernel': 'sigmoid'}	ngram	0.74747475	0.74070194	0.74747475
SVM	{'classifier__C': 1, 'classifier__kernel': 'sigmoid'}	cluster	0.74747475	0.73884141	0.74747475
Naive Bayes	{}	cluster	0.74747475	0.74747475	0.74747475
Logistic regression	{}	word2vec	0.74747475	0.74737164	0.74747475
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 10}	demographics	0.74489796	0.73877812	0.74489796
Logistic regression	{}	ngram	0.73737374	0.71626984	0.73737374
Logistic regression	{}	ngram	0.73469388	0.70726103	0.73469388
Logistic regression	{}	cluster	0.72727273	0.72320596	0.72727273
Naive Bayes	{}	ngram	0.7244898	0.72214638	0.7244898
SVM	{'classifier__C': 5, 'classifier__kernel': 'linear'}	word2vec	0.71717172	0.71714286	0.71717172
KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 10}	word2vec	0.71717172	0.71714286	0.71717172
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 10}	word2vec	0.71717172	0.71714286	0.71717172
XGBoost	{'classifier__max_depth': 6, 'classifier__n_estimators': 50}	word2vec	0.71428571	0.71236897	0.71428571
SVM	{'classifier__C': 1, 'classifier__kernel': 'rbf'}	demographics	0.71428571	0.71416667	0.71428571
SVM	{'classifier__C': 5, 'classifier__kernel': 'rbf'}	demographics	0.71428571	0.71416667	0.71428571
SVM	{'classifier__C': 10, 'classifier__kernel': 'rbf'}	demographics	0.71428571	0.71416667	0.71428571
Decision tree	{}	demographics	0.71428571	0.71416667	0.71428571
XGBoost	{'classifier__max_depth': 6, 'classifier__n_estimators': 50}	demographics	0.71428571	0.71416667	0.71428571
XGBoost	{'classifier__max_depth': 10, 'classifier__n_estimators': 50}	demographics	0.71428571	0.71416667	0.71428571
Logistic regression	{}	demographics	0.71428571	0.71416667	0.71428571
Decision tree	{}	cluster	0.70707071	0.70514532	0.70707071
Naive Bayes	{}	cluster	0.70707071	0.70109318	0.70707071
Decision tree	{}	ngram	0.70408163	0.7025641	0.70408163
Logistic regression	{}	cluster	0.70408163	0.70156463	0.70408163

Logistic regression	{}	demographics	0.6969697	0.69619476	0.6969697
SVM	{'classifier__C': 5, 'classifier__kernel': 'linear'}	word2vec	0.68686869	0.68635667	0.68686869
Decision tree	{}	cluster	0.68367347	0.67800742	0.68367347
Decision tree	{}	ngram	0.67676768	0.67673469	0.67676768
Naive Bayes	{}	word2vec	0.67676768	0.66917293	0.67676768
SVM	{'classifier__C': 5, 'classifier__kernel': 'sigmoid'}	cluster	0.67346939	0.66849894	0.67346939
XGBoost	{'classifier__max_depth': 10, 'classifier__n_estimators': 10}	word2vec	0.66666667	0.66543779	0.66666667
Decision tree	{}	word2vec	0.65306122	0.65175585	0.65306122
Decision tree	{}	ngram	0.64646465	0.64646465	0.64646465
Logistic regression	{}	word2vec	0.64646465	0.64646465	0.64646465
SVM	{'classifier__C': 1, 'classifier__kernel': 'rbf'}	demographics	0.64646465	0.63437797	0.64646465
SVM	{'classifier__C': 5, 'classifier__kernel': 'rbf'}	demographics	0.64646465	0.63437797	0.64646465
SVM	{'classifier__C': 10, 'classifier__kernel': 'rbf'}	demographics	0.64646465	0.63437797	0.64646465
Decision tree	{}	demographics	0.64646465	0.63437797	0.64646465
XGBoost	{'classifier__max_depth': 6, 'classifier__n_estimators': 10}	demographics	0.64646465	0.63437797	0.64646465
XGBoost	{'classifier__max_depth': 6, 'classifier__n_estimators': 50}	demographics	0.64646465	0.63437797	0.64646465
XGBoost	{'classifier__max_depth': 10, 'classifier__n_estimators': 10}	demographics	0.64646465	0.63437797	0.64646465
XGBoost	{'classifier__max_depth': 10, 'classifier__n_estimators': 50}	demographics	0.64646465	0.63437797	0.64646465
KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 10}	demographics	0.64646465	0.63437797	0.64646465
SVM	{'classifier__C': 1, 'classifier__kernel': 'rbf'}	fall_location	0.63636364	0.62998339	0.63636364
KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 10}	word2vec	0.63265306	0.63204005	0.63265306
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 10}	word2vec	0.63265306	0.63204005	0.63265306
SVM	{'classifier__C': 1, 'classifier__kernel': 'sigmoid'}	fall_location	0.62626263	0.6068477	0.62626263
Decision tree	{}	word2vec	0.62626263	0.61863613	0.62626263
KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 5}	fall_location	0.62244898	0.61763155	0.62244898
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 5}	fall_location	0.62244898	0.61763155	0.62244898
Logistic regression	{}	fall_location	0.61616162	0.60714286	0.61616162
Decision tree	{}	fall_location	0.61616162	0.61580882	0.61616162
Logistic regression	{}	fall_location	0.61616162	0.60449958	0.61616162
Naive Bayes	{}	word2vec	0.6122449	0.60964361	0.6122449

KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 5}	fall_location	0.60606061	0.59802186	0.60606061
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 5}	fall_location	0.60606061	0.59802186	0.60606061
Naive Bayes	{}	cluster	0.60204082	0.59696299	0.60204082
Decision tree	{}	fall_location	0.5959596	0.59558824	0.5959596
Naive Bayes	{}	fall_location	0.58585859	0.48391608	0.58585859
XGBoost	{'classifier__max_depth': 6, 'classifier__n_estimators': 10}	fall_location	0.57575758	0.55576923	0.57575758
XGBoost	{'classifier__max_depth': 6, 'classifier__n_estimators': 50}	fall_location	0.57575758	0.55576923	0.57575758
XGBoost	{'classifier__max_depth': 10, 'classifier__n_estimators': 10}	fall_location	0.57575758	0.55576923	0.57575758
XGBoost	{'classifier__max_depth': 10, 'classifier__n_estimators': 50}	fall_location	0.57575758	0.55576923	0.57575758
Naive Bayes	{}	fall_location	0.57575758	0.49782609	0.57575758
KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 5}	cluster	0.57575758	0.48793103	0.57575758
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 5}	cluster	0.57575758	0.48793103	0.57575758
XGBoost	{'classifier__max_depth': 6, 'classifier__n_estimators': 10}	fall_location	0.57575758	0.54166667	0.57575758
XGBoost	{'classifier__max_depth': 6, 'classifier__n_estimators': 50}	fall_location	0.57575758	0.54166667	0.57575758
XGBoost	{'classifier__max_depth': 10, 'classifier__n_estimators': 10}	fall_location	0.57575758	0.54166667	0.57575758
XGBoost	{'classifier__max_depth': 10, 'classifier__n_estimators': 50}	fall_location	0.57575758	0.54166667	0.57575758
KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 5}	cluster	0.56122449	0.45660864	0.56122449
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 5}	cluster	0.56122449	0.45660864	0.56122449
Decision tree	{}	cluster	0.55555556	0.55514706	0.55555556
KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 50}	word2vec	0.54545455	0.53867661	0.54545455
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 50}	word2vec	0.54545455	0.53867661	0.54545455
KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 5}	fall_location	0.54545455	0.50593324	0.54545455
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 5}	fall_location	0.54545455	0.50593324	0.54545455
Naive Bayes	{}	demographics	0.54081633	0.44339265	0.54081633
KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 5}	cluster	0.53535354	0.3675	0.53535354
KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 10}	cluster	0.53535354	0.3675	0.53535354
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 5}	cluster	0.53535354	0.3675	0.53535354
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 10}	cluster	0.53535354	0.3675	0.53535354
Naive Bayes	{}	word2vec	0.53535354	0.52954545	0.53535354
Naive Bayes	{}	fall_location	0.53061224	0.5006646	0.53061224

KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 5}	ngram	0.52525253	0.34437086	0.52525253
KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 10}	ngram	0.52525253	0.34437086	0.52525253
KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 50}	ngram	0.52525253	0.34437086	0.52525253
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 5}	ngram	0.52525253	0.34437086	0.52525253
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 10}	ngram	0.52525253	0.34437086	0.52525253
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 50}	ngram	0.52525253	0.34437086	0.52525253
Naive Bayes	{}	demographics	0.52525253	0.42105263	0.52525253
KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 5}	ngram	0.52525253	0.34437086	0.52525253
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 5}	ngram	0.52525253	0.34437086	0.52525253
KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 5}	ngram	0.52040816	0.34228188	0.52040816
KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 10}	ngram	0.52040816	0.34228188	0.52040816
KNN	{'classifier__leaf_size': 30, 'classifier__n_neighbors': 50}	ngram	0.52040816	0.34228188	0.52040816
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 5}	ngram	0.52040816	0.34228188	0.52040816
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 10}	ngram	0.52040816	0.34228188	0.52040816
KNN	{'classifier__leaf_size': 50, 'classifier__n_neighbors': 50}	ngram	0.52040816	0.34228188	0.52040816
SVM	{'classifier__C': 1, 'classifier__kernel': 'sigmoid'}	fall_location	0.51020408	0.49679076	0.51020408
SVM	{'classifier__C': 5, 'classifier__kernel': 'sigmoid'}	fall_location	0.51020408	0.49679076	0.51020408
SVM	{'classifier__C': 10, 'classifier__kernel': 'sigmoid'}	fall_location	0.51020408	0.49679076	0.51020408
Decision tree	{}	fall_location	0.51020408	0.49679076	0.51020408
Logistic regression	{}	fall_location	0.51020408	0.49679076	0.51020408
XGBoost	{'classifier__max_depth': 6, 'classifier__n_estimators': 10}	fall_location	0.5	0.47593583	0.5
XGBoost	{'classifier__max_depth': 6, 'classifier__n_estimators': 50}	fall_location	0.5	0.47593583	0.5
XGBoost	{'classifier__max_depth': 10, 'classifier__n_estimators': 10}	fall_location	0.5	0.47593583	0.5
XGBoost	{'classifier__max_depth': 10, 'classifier__n_estimators': 50}	fall_location	0.5	0.47593583	0.5
XGBoost	{'classifier__max_depth': 10, 'classifier__n_estimators': 50}	word2vec	0.49494949	0.49489796	0.49494949
Naive Bayes	{}	demographics	0.49494949	0.36342593	0.49494949
Decision tree	{}	word2vec	0.45454545	0.4544898	0.45454545