Machine Learning - Final Project

Parsa Jafaripour and Robbie Bernstein

Datasets

Binary Dataset - ISOT Fake News Dataset

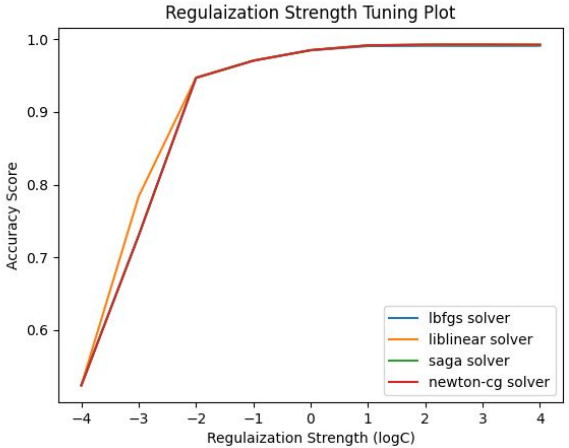
- 44,919 examples
- 4 features: title, text, subject, date
- Models to train: logistic regression, Naive Bayes, random forest
- Metrics: accuracy and recall

Multiclass Dataset - Customer Segmentation Dataset

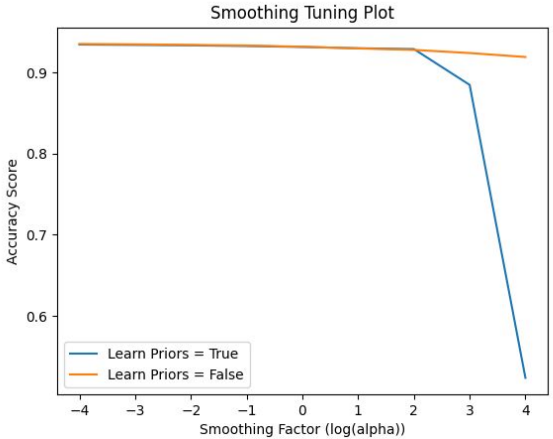
- 2,627 examples
- 9 features: gender, marital status, age, graduation status, profession, work experience, spending score, family size, Var_1
- Models to train: random forest, support vector machine, artificial neural network
- Metrics: accuracy and F1 score

Both datasets from Kaggle

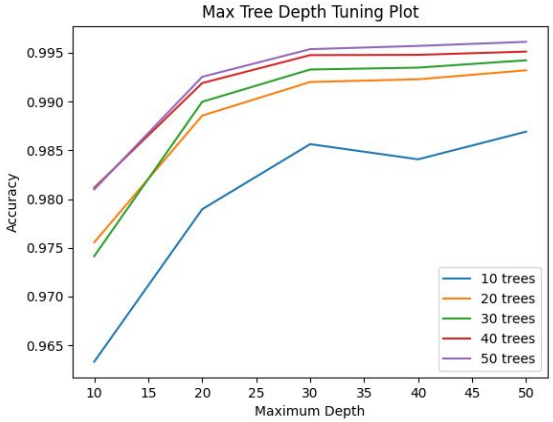
Hyperparameter Tuning - Binary Dataset



Logistic Regression

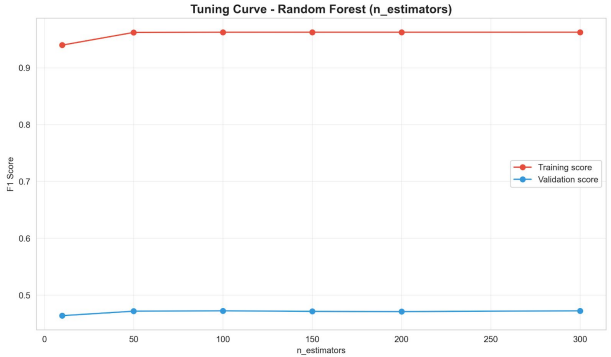


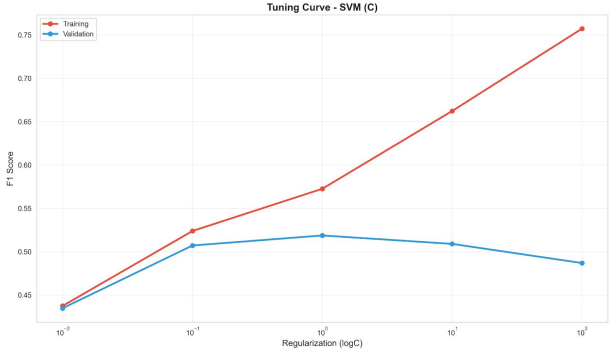
Naive Bayes

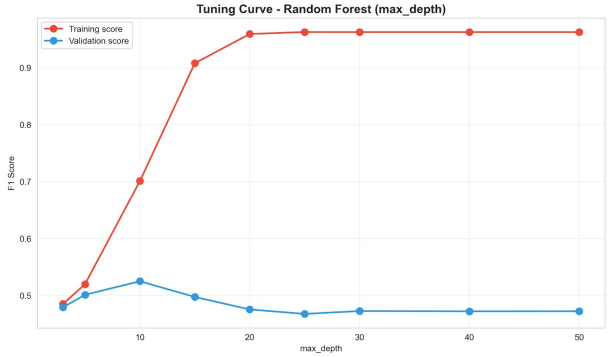


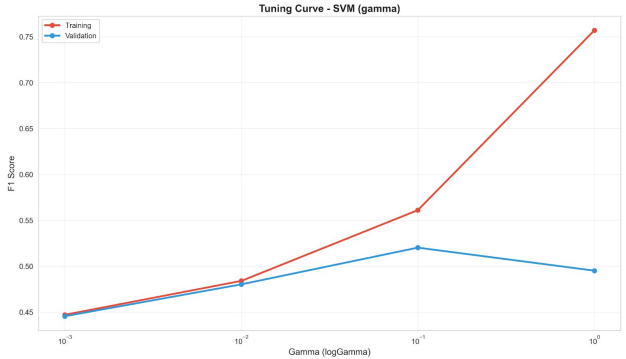
Random Forest

Hyperparameter Tuning - Multiclass Dataset

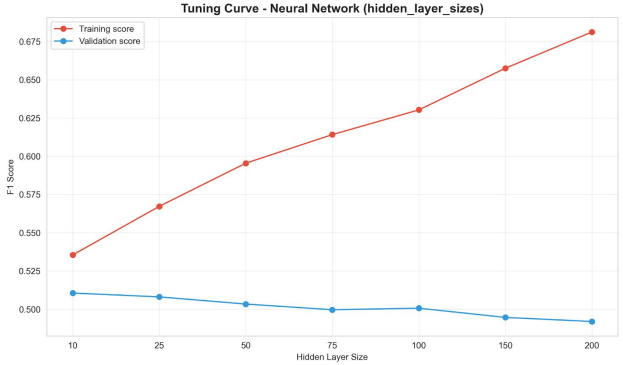


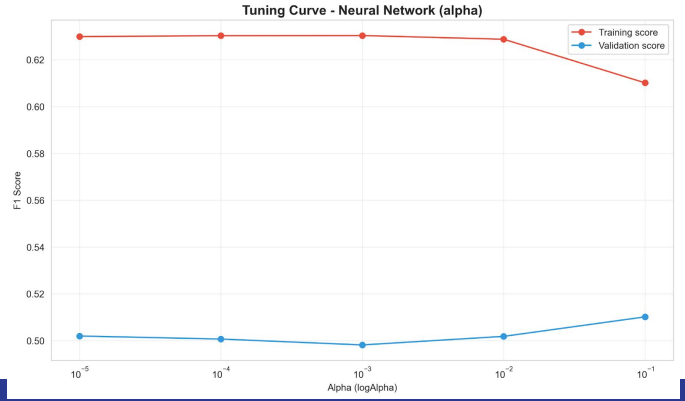


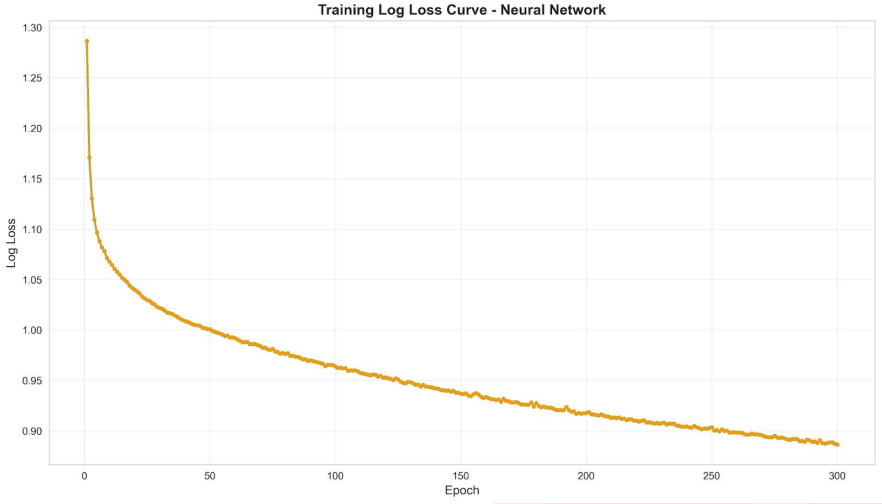




Hyperparameter Tuning - Multiclass Dataset

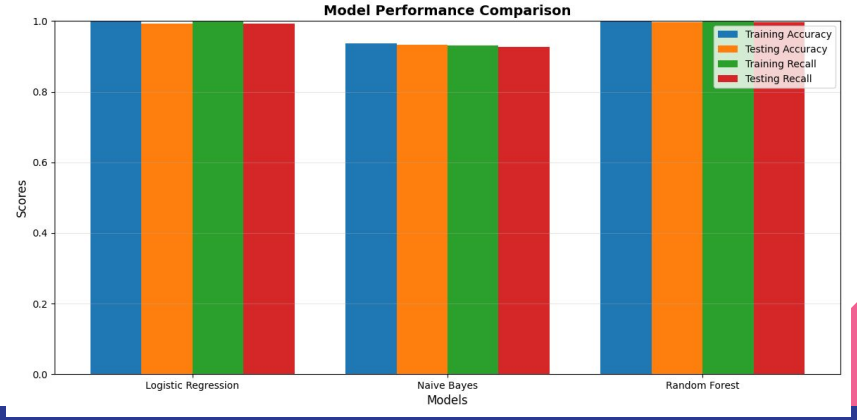






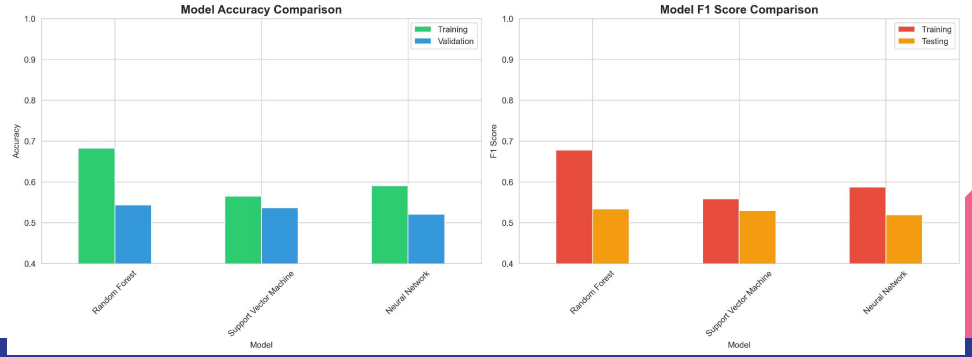
Results - Binary Dataset

Model	Hyperparameters	Training Accuracy	Testing Accuracy	Training Recall	Testing Recall
Logistic Regression	C = 1000 solver = newton-cg	1.0000	0.9931	0.9999	0.9926
Naive Bayes	alpha = 0.0001 fit priors = False	0.9374	0.9323	0.9308	0.9260
Random Forest	# trees = 50 max depth = None	1.0000	0.9960	0.9999	0.9963



Results - Multiclass Dataset

Model	Hyperparameters	Training Accuracy	Testing Accuracy	Training F1 Score	Testing F1 Score
Random Forest	# trees = 100 max depth = 10	0.6828	0.5434	0.6778	0.5338
Support Vector Machine (RBF Kernel)	C = 1 gamma = 0.1	0.5648	0.5366	0.5582	0.5299
Artificial Neural Network	alpha = 0.001 hidden layers = 50	0.5905	0.5211	0.5872	0.5186



Conclusion

Fake News Dataset

- Highly learnable
- Large dataset provides ample training data
- Likely little or no noise, clear separation between classes

Customer Segmentation Dataset

- Difficult to learn
- Likely a large amount of noise and class overlap
- Small dataset might limit models’ ability to learn class boundaries