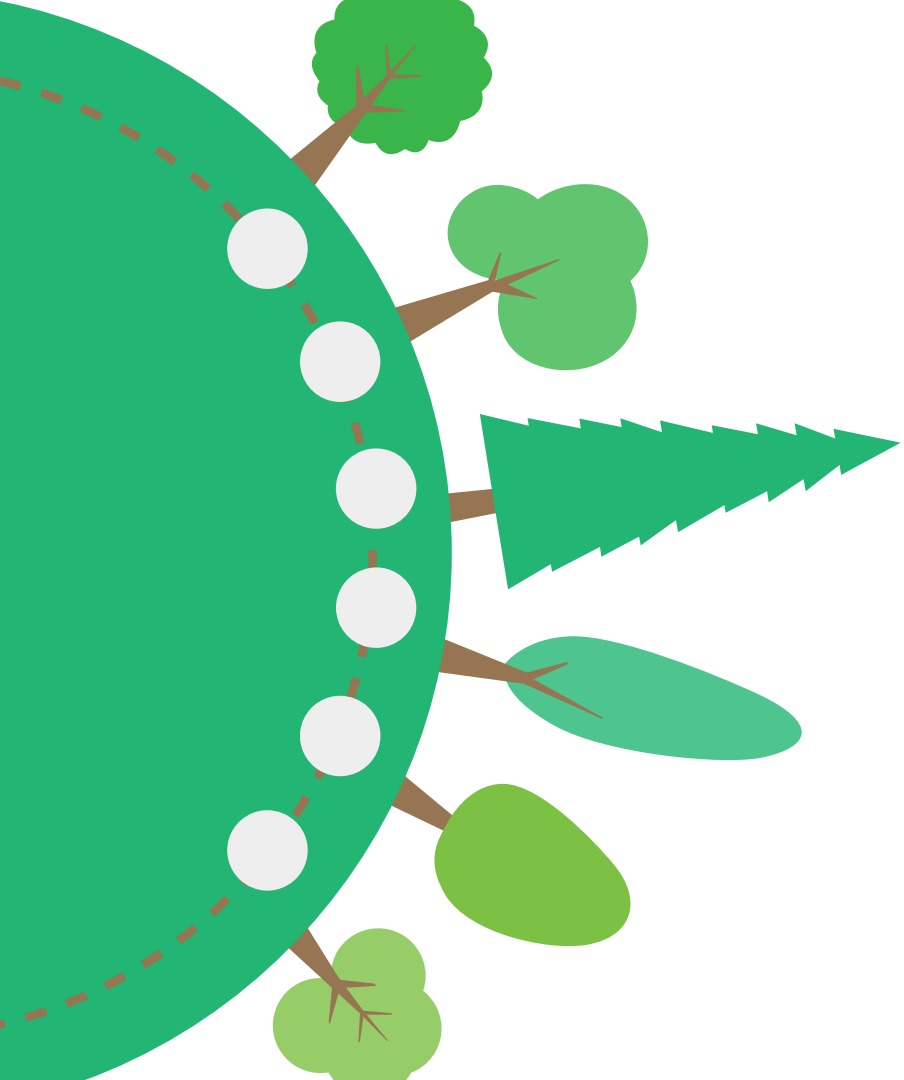


Hackathon

-Random Forest-

algorithm constraint

Nolan Arendt, Jessica Bow, and Peter Yonka

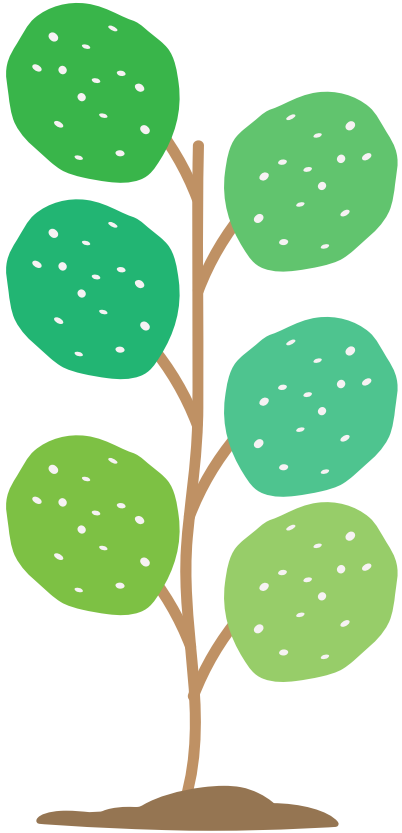
Problem Statement

Given the constraint of only using a Random Forest Model, we endeavor to accurately predict if an individual makes greater than or less than $50,000 per year.

Cleaning & Prep

1. -Dropped rows where working class and occupation were '?'
2. -Realized those with '?' in column Occupation had never worked.
   -Replaced '?' in Native Country with Unknown
3. -Binarized Sex and Wage columns.
   -Dummified remaining categorical columns.
4. - Using correlation table, chose top 10 positively and negatively correlated features
5. - Got a baseline accuracy score of 75.1%

Model Scoring Summary



Model	Feature Set	Best Parameters	Model Accuracy Score	Training Accuracy Score	Testing Accuracy Score	Improv. Over Baseline
Random Forest	Limited*	{max_depth: 12, n_estimators: 90}	84.9%	86.6%	85.4%	+10.3%
	Full	{max_depth: 15, n_estimators: 75}	85.8%	88.5%	85.9%	+10.8%
AdaBoost using Random Forest	Limited*	{rf_max_depth: 4, learning_rate: 0.90, n_estimators: 90}	85.3%	86.2%	85.6%	+10.5%
	Full	**{rf_max_depth: 4, learning_rate: 0.90, n_estimators: 90}**	**86.7%**	**88.6%**	**86.7%**	**+11.6%**

* Top 20 features with strongest positive and negative correlations

Future Analysis