Feature selection for Predictive Modeling

Karthik Prakash
kprakas2@stanford.edu

Motivation

- Predictive models are trained on a scheduled basis on the recent dataset.
- **WHY WOULD SOMEONE TOUCH CODE THAT IS WORKING IN PRODUCTION!!!**
- Lots of things change in few months in Industry. The features you thought was the most important might be obsolete because the data changed.
- You are **BURNING money** hosting these obsolete features. For real-time model predictions you need the feature values in super-fast database which is not CHEAP.
- **Goal**: **Build a framework that gives meaningful insights of the features which will aid in maintaining feature data quality.**

Data

CONSTRAINTS WHILE CHOOSING DATA

- A simple dataset with few features so that it is easy to do feature selection experiments.
- NO correlated features.
- Synthetically generate data to vary dataset size and features.

SYNTHETIC DATA GENERATION

- All features were univariate normal distributions separately.
- Sample from Normal Distribution for each feature to produce large datasets with same distribution.
- Verified Data distribution of synthetic and original data are identical.
- No change in the model accuracy for synthetic data compared to original data.

SYNTHETIC Features Added

- Random features.
- Redundant features
- Low Variance features
- Zero Variance features

Technique

Recursive Feature Elimination

```
X = training data
for n in range(number_of_features):
    θ = train_logistic_regression(X)
    least_important feature = min value in θ
    X.drop(least important feature)
    return feature ranks
```

- Most **effective** way of selecting feature if cross validated with original model
- Effective for **redundant features, low variance features, random features**.
- **EXPENSIVE**: N! number of training jobs.

Variance Threshold Elimination

- All the features below certain variance threshold are eliminated.

- Easiest way to identify "aging" features losing importance over a period of time.
- Not effective on any other type of irrelevant features.

Chi Square Statistics

- chi-squared stats between each non-negative feature and class.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- Good to find low variance features and random features
- Not effective against redundant features
- Not effective when feature values have negative numbers

Tree Based Feature Selection

- ExtraTreeClassifier trained on the data.
- Feature importance scores available from the trained tree classifier.

- Regression Models not supported
- Not effective against redundant feature
- Effective against Low variance and ransom features.

Analysis

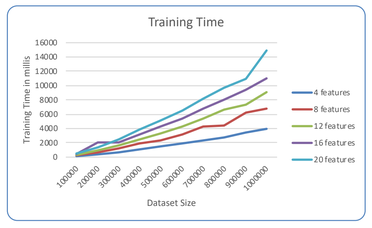
Future

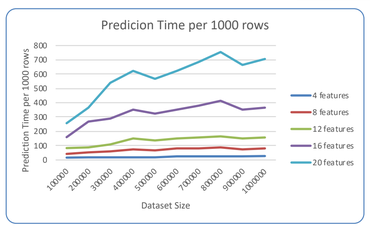
- I would make this an open source framework with self-serve usage documentation.
- Automatic Model training on different feature selection datasets and comparison in model accuracy, prediction time and training time if a given feature set is selected or not.

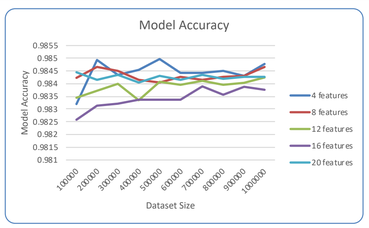
Feature Selection Report

	Technique Name	sepal_length	sepal_width	petal_length	petal_width	0_var_feature_1	low_var_feature_1	redundant_feature	random_feature
0	Recursive Feature Elimination Rank	5	1	1	1	1	4	2	3
1	Zero Variance Elimination	True	True	True	True	False	True	True	True
2	Low Variance(0.1) Elimination	True	True	True	True	False	False	True	False
3	2 Best features based on Chi Square Statistics	False	False	True	True	False	False	False	False
4	2 Best features based on Anova F Values	False	False	True	True	False	False	False	False
5	3 Best features based on Chi Square Statistics	False	False	True	True	False	False	True	False
6	3 Best features based on Anova F Values	False	False	True	True	False	False	True	False
7	Redundant Features	redundant_feature						sepal_length	
8	Feature Importance scores	0.08409302221488237	0.08428527412897678	0.31824997364834545	0.38636452852331743	0.0	0.014480139566432251	0.09614643735564886	0.016388624562396983

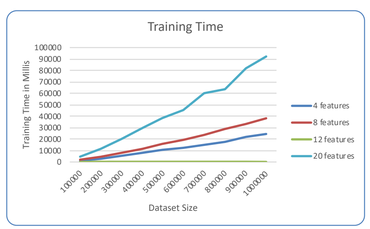
Results

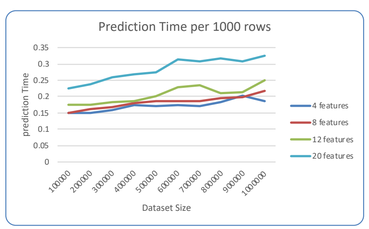


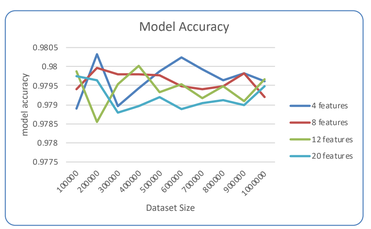




- KNN Classifier Models trained on different datasets sizes and different feature counts
- Dataset Size has high impact on training time
- irrelevant feature count has impact on training time
- irrelevant feature count has impact on prediction time







- Decision Tree classifier trained on different dataset sizes and different feature counts
- Dataset Size has high impact on training time
- irrelevant feature count has impact on training time
- irrelevant feature count has impact on prediction time