ANALYZING DONATIONS TO 2016 PRESIDENTIAL CANDIDATES

Raphael Palefsky-Smith and Christina Wadsworth **/** CS 229: Machine Learning **/** December 8, 2015

The problem

Given just a donor's…

- name
- address
- job

can we predict **Democrat** or **Republican**?

Humans get it right only **50%** of the time.

The data

Mandatory disclosures from Hillary Clinton, Bernie Sanders, Jeb Bush, and Ben Carson. Over **250,000** donations with over **100,000** unique donors.

Campaign	Contributor type	Last Name	First Name
BERNIE 2016	IND (individual)	[redacted]	[redacted]
HILLARY FOR AMERICA	IND (individual)	[redacted]	[redacted]
Address	**City**	**State**	**Zip Code**
[redacted]	Palo Alto	CA	94306-1518
[redacted]	Stanford	CA	94305-1068
Contribution Date	**Contribution Amount**	**Employer**	**Job Title**
2015-06-25	$10.00	Stanford University	Professor
2015-09-18	$100.00	Stanford University	Professor

Augmented with **average household income by ZIP code** statistics from the IRS.

Zipcode	Total Adjusted Gross Income	# Returns
33109	435,729,000	240

Pre-processing

Remove **donation date** - while it reduces error on the test set, it means the model won't generalize.

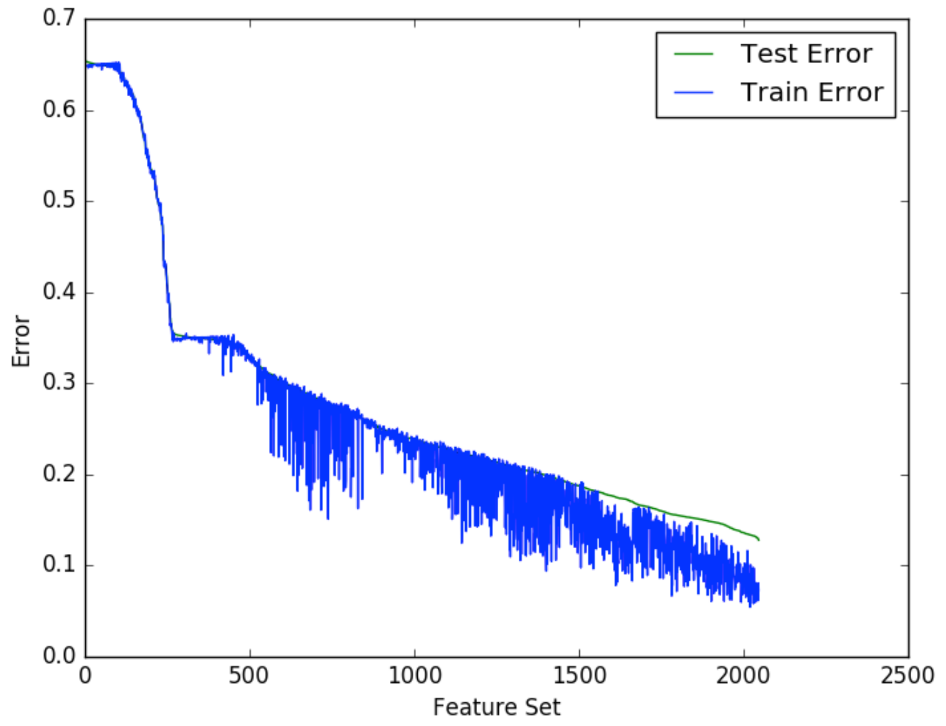
Remove **amount** to further generalize.

Remove **title** (Mr./Mrs.) - it helped reduce our error, but this was misleading. We discovered that only Republican candidates populate this field!

Correlate **zip code** with IRS data to find average income in donor's area. Additionally, discretize this feature into **tax bracket** to remove outliers and contextualize data.

Analysis

Stochastic gradient descent with an **SVM** objective function. Since there were only 2047 possible subsets of features (and analysis is embarassingly parallel), feature selection done via **brute-force.**

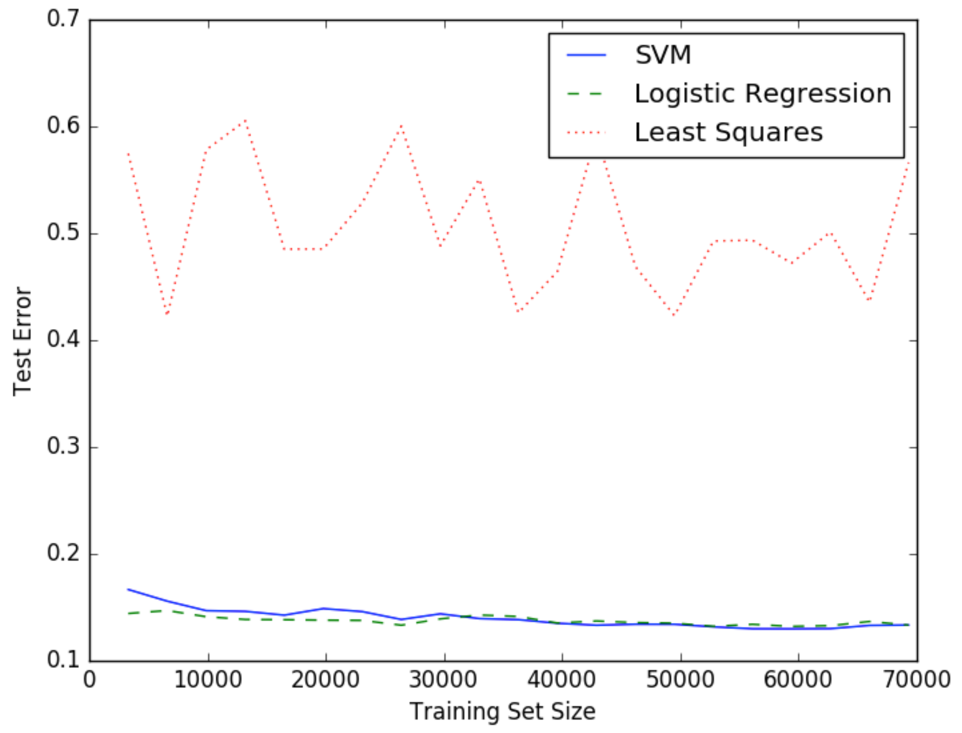


Best feature set (12% error): first/middle/last name, employer, occupation, state, zip code, tax bracket
Worse feature set (65% error): area income, city, tax bracket

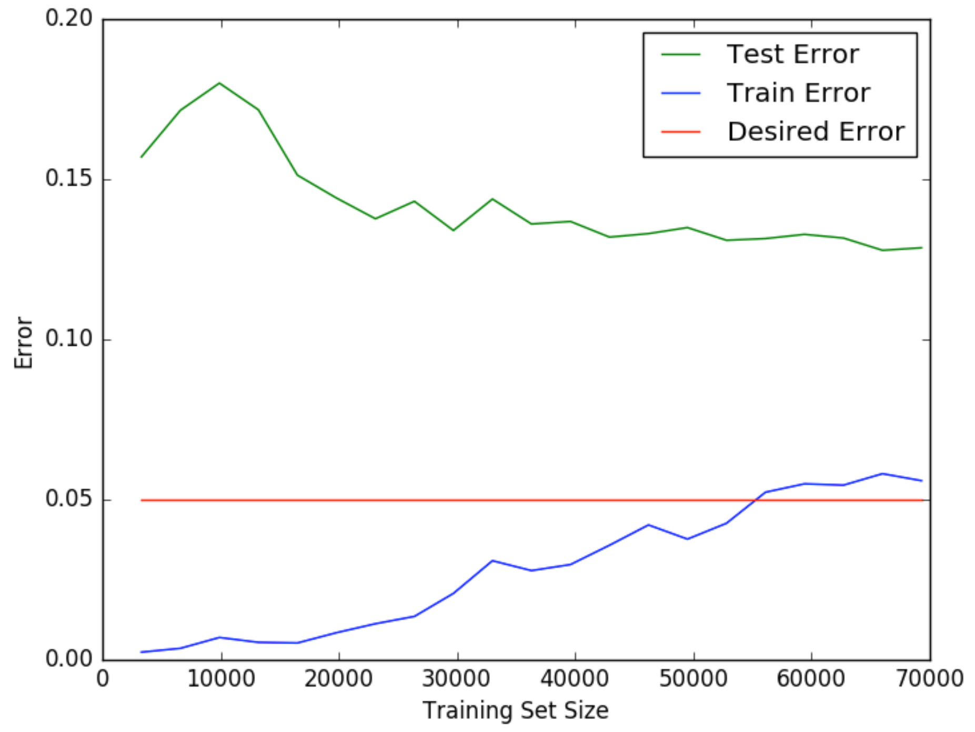
Results

87.4% SVM accuracy / **90.5%** SVM F1

Comparison of learning models:



Error curve (SVM):



Confusion matrix (SVM):

	Republican	Democrat
Republican	9258	2891
Democrat	1565	21239

Insights:

- model still suffers from **high variance**
- adding income data doesn't significantly reduce error
- specific algorithm doesn't matter much