Clustering, Regression, & Classification Using Census Data

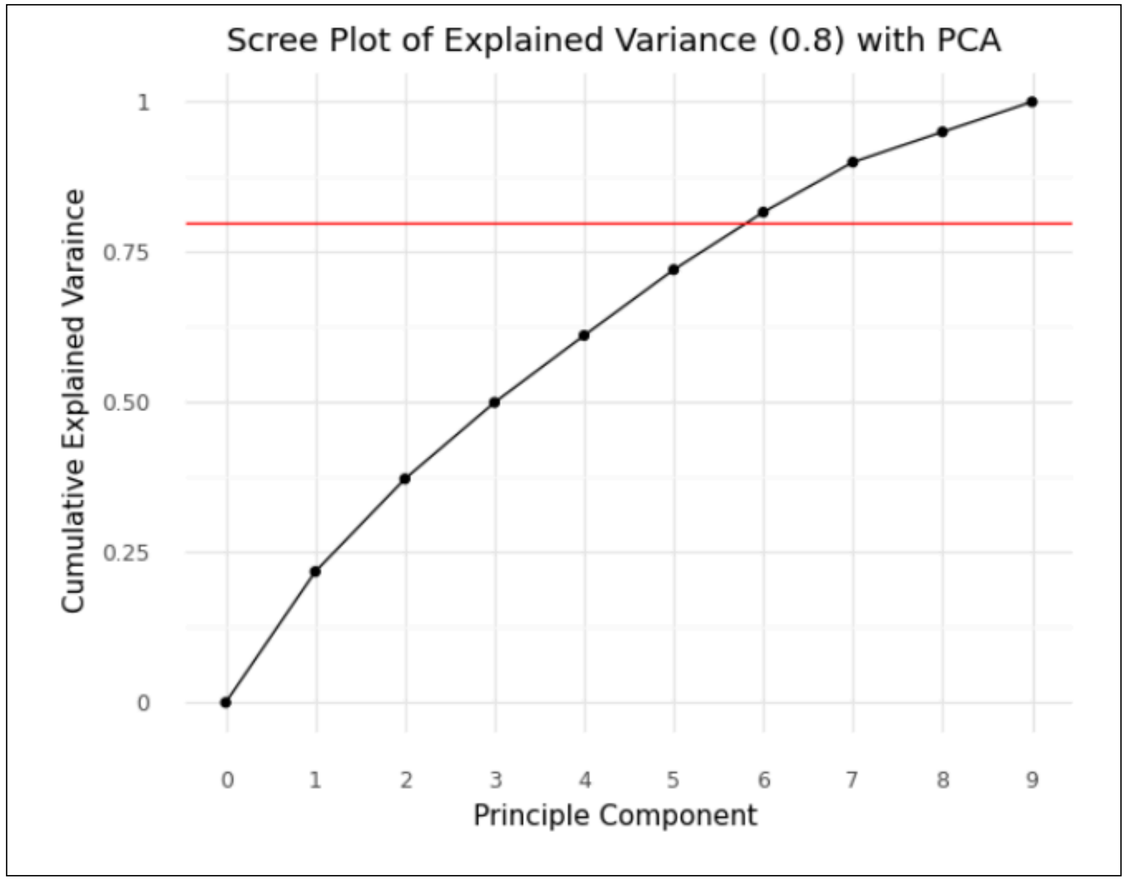
CPSC392-01

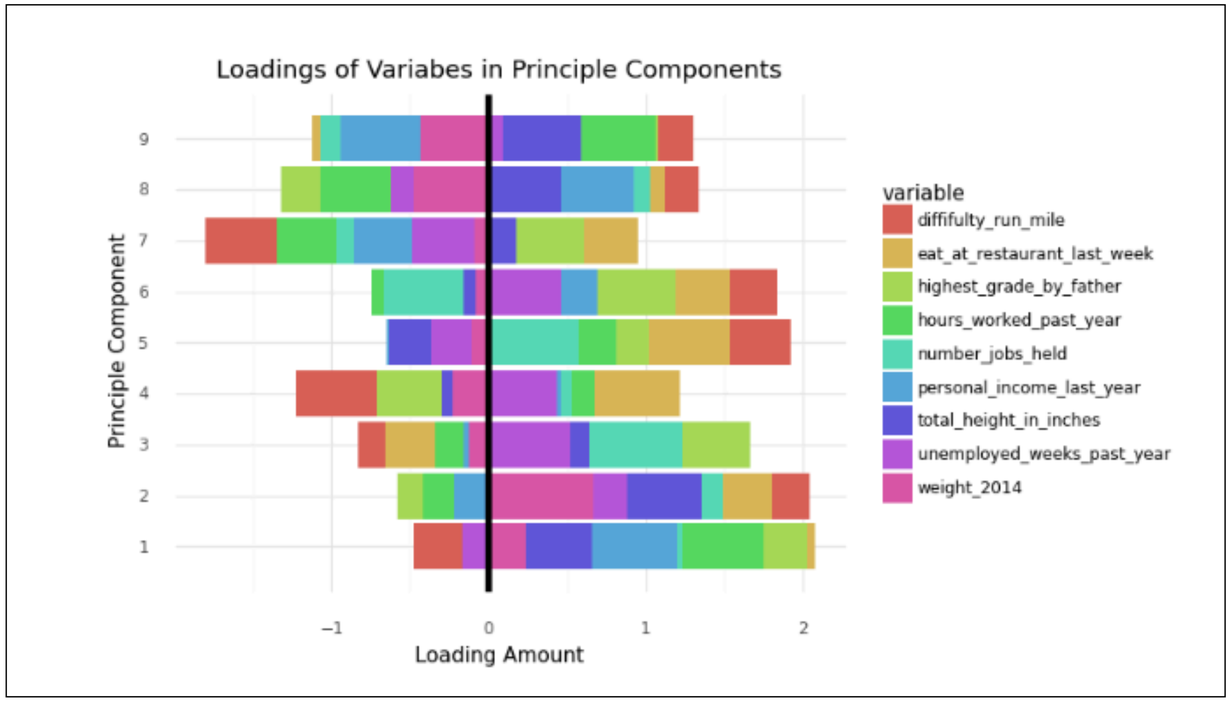
Connor Lydon 12/13/2021

Three Questions

1. What groups exist within the data?
2. How well can a model predict Annual Income?
3. Can you predict who has gone bankrupt?

Principle Component Analysis

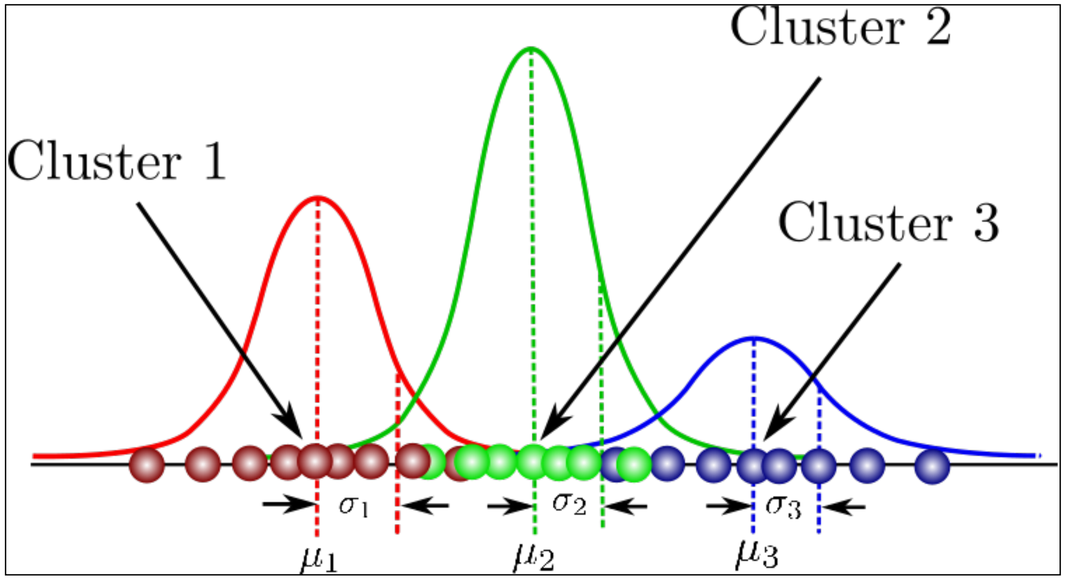




1. What groups exist within the data?

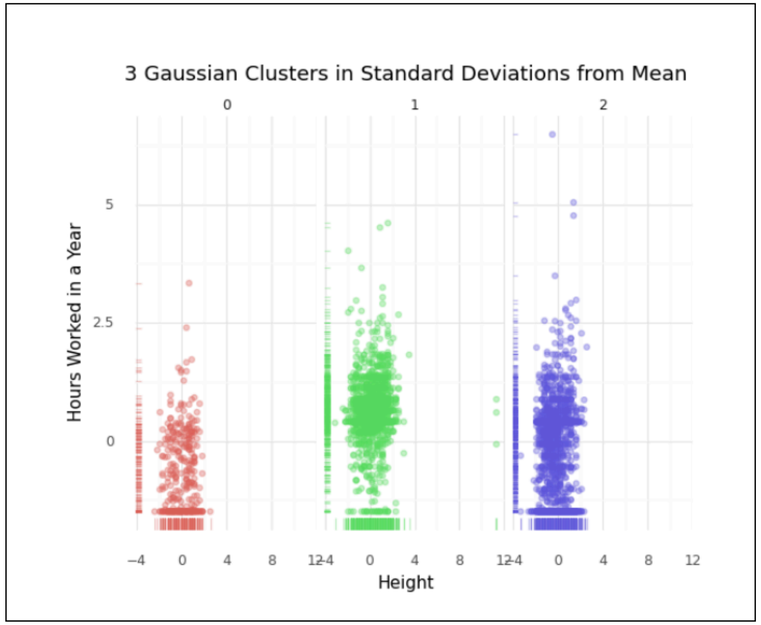
Evaluating Clustering Algorithms

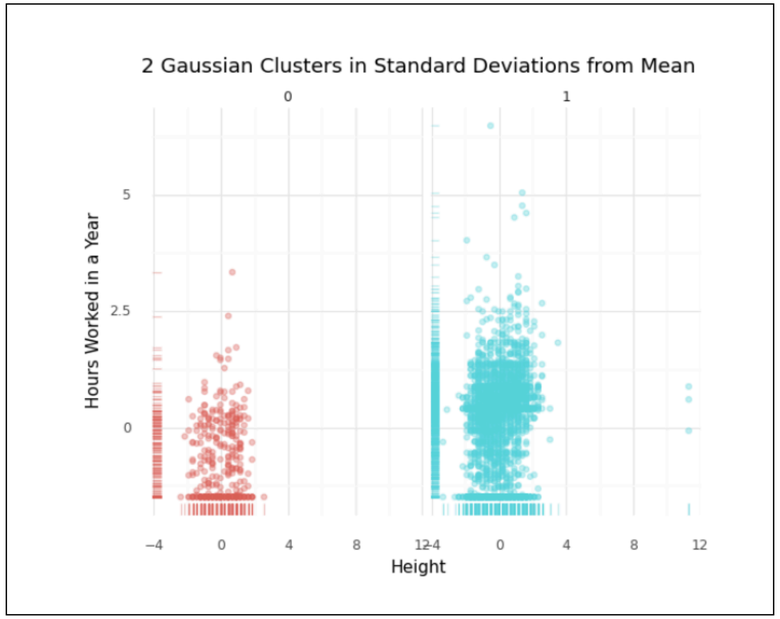
```
Loaded Q1 Data
all_vars
em score for 2:  0.25899572996071596
hier score for 2:  0.12427601935633445
em score for 3:  0.25647046468889595
hier score for 3:  0.1379959489599073
em score for 4:  0.08047743989908811
hier score for 4:  0.10685550537272899
em score for 5:  0.06644074762525108
hier score for 5:  0.12210473793626674

financial
physical
em score for 2:  0.36601929612052986
em score for 2:  0.22200020552104244
em score for 3:  0.12600317326219332
em score for 3:  0.06998094702961187
em score for 4:  0.1995150543228401
em score for 4:  0.2089796446153722
```

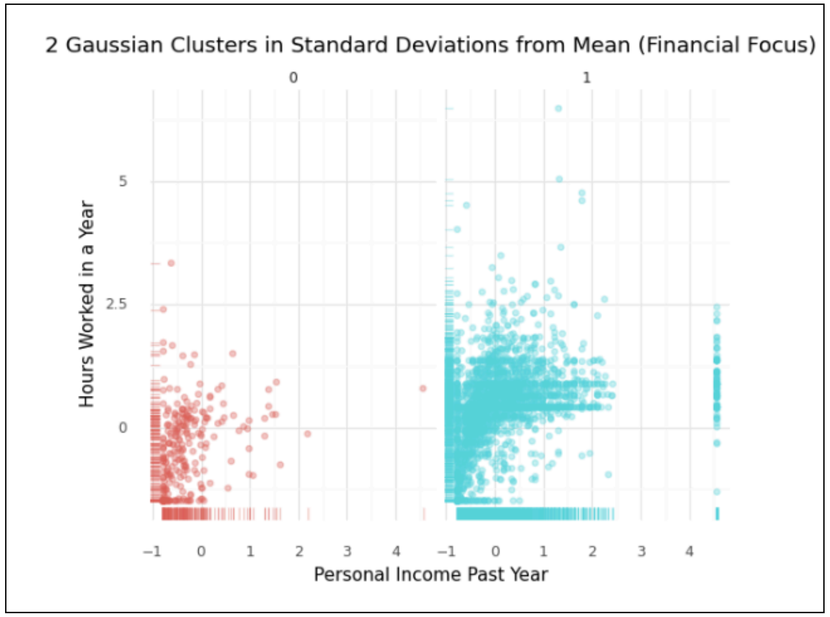


Towards Data Science Oscar Contreras Carrasco

Gaussian Clusters







1. What groups exist within the data?

Mean Values For Clusters

Gaussian 3 Clustered with All Continuous Variables

em_3_cluster	highest_grade_by_father	personal_income_last_year	total_height_in_inches	declared_bankruptcy	missed_payment_on_loan	male	size
0	-0.119589	-0.410408	0.039009	0.184211	0.248538	0.543860	342
1	0.216290	0.635402	0.300653	0.149924	0.120857	0.628251	1961
2	-0.163014	-0.470295	-0.256453	0.207146	0.163760	0.350489	2351

Gaussian 2 Clustered with All Continuous Variables

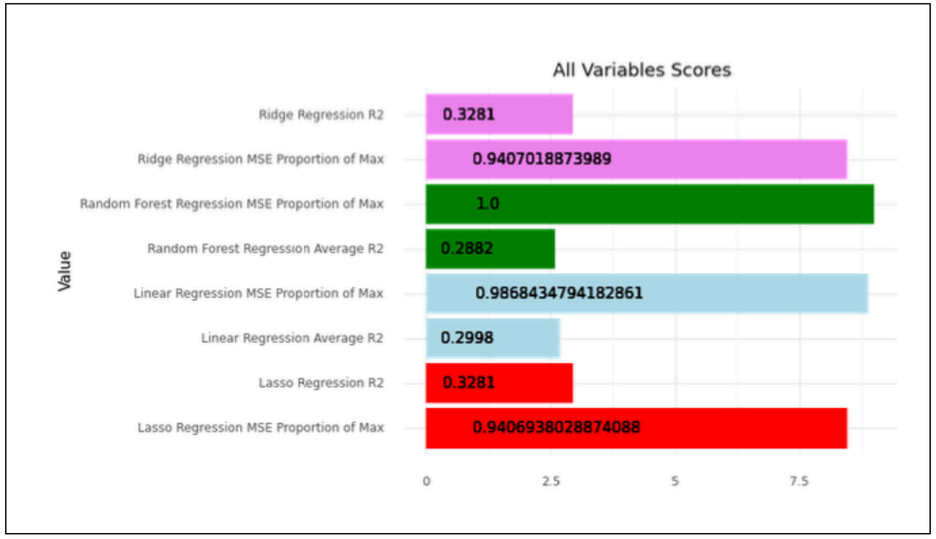
em_2_cluster	highest_grade_by_father	personal_income_last_year	total_height_in_inches	declared_bankruptcy	missed_payment_on_loan	male	size
0	-0.119589	-0.410408	0.039009	0.184211	0.248538	0.543860	342
1	0.009485	0.032551	-0.003094	0.181122	0.144249	0.476809	4312

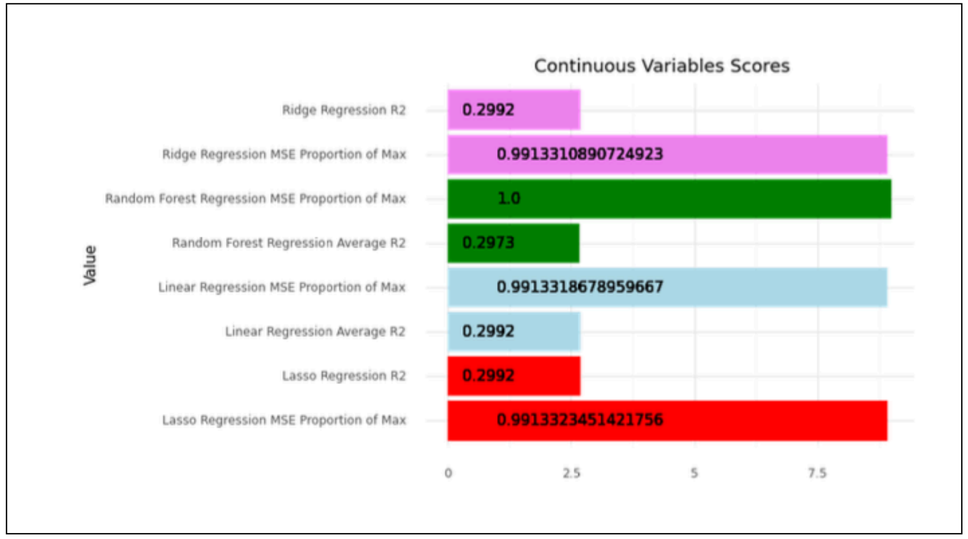
Gaussian 2 Clustered with All Continuous Financial Variables

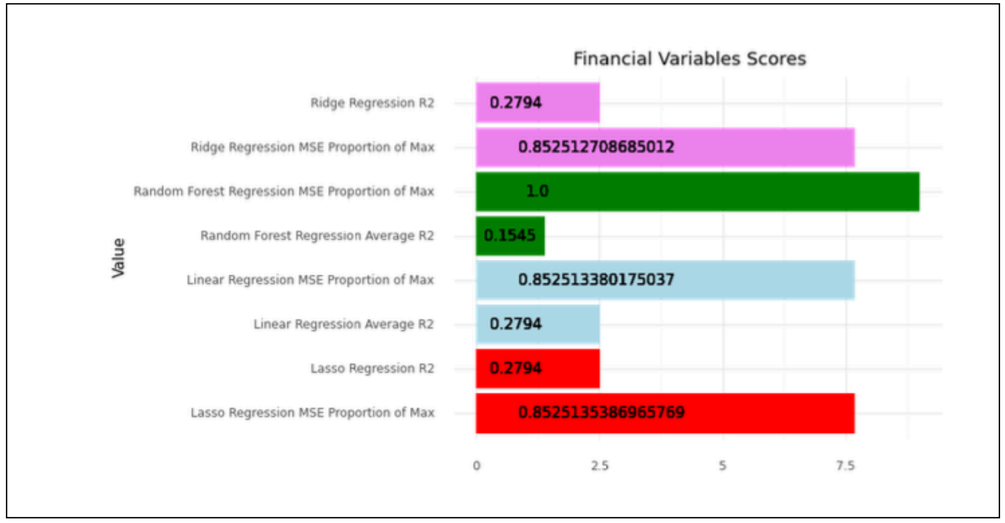
em_2_cluster_financial	highest_grade_by_father	personal_income_last_year	total_height_in_inches	declared_bankruptcy	missed_payment_on_loan	male	size
0	-0.119589	-0.410408	0.039009	0.184211	0.248538	0.543860	342
1	0.009485	0.032551	-0.003094	0.181122	0.144249	0.476809	4312

1. What groups exist within the data?

Regression Model Scores

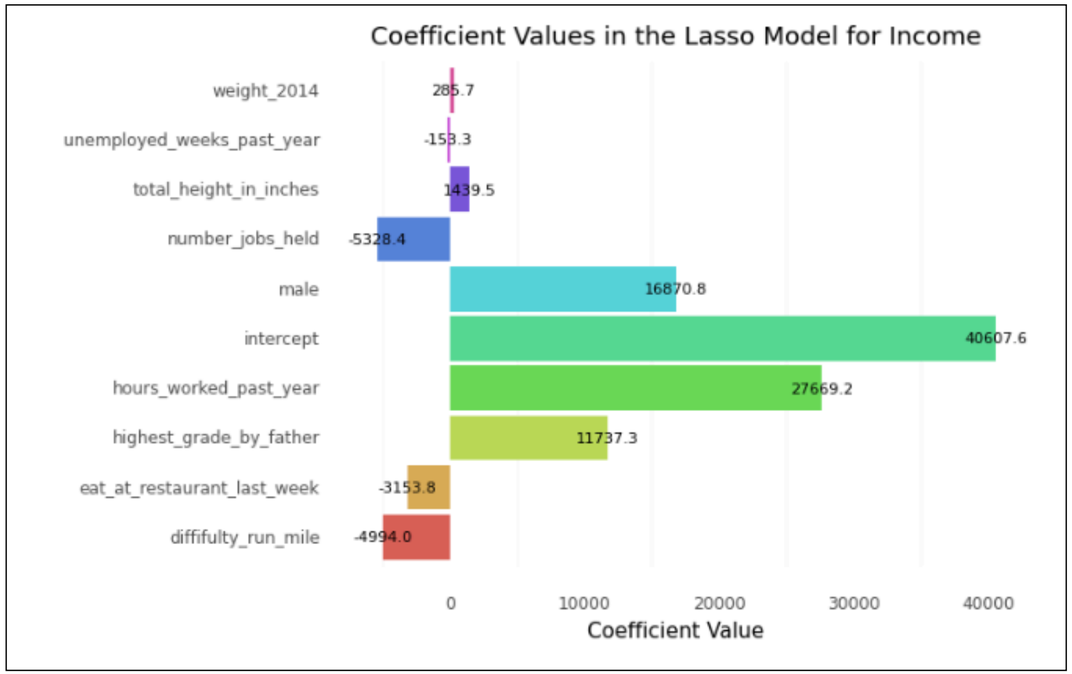


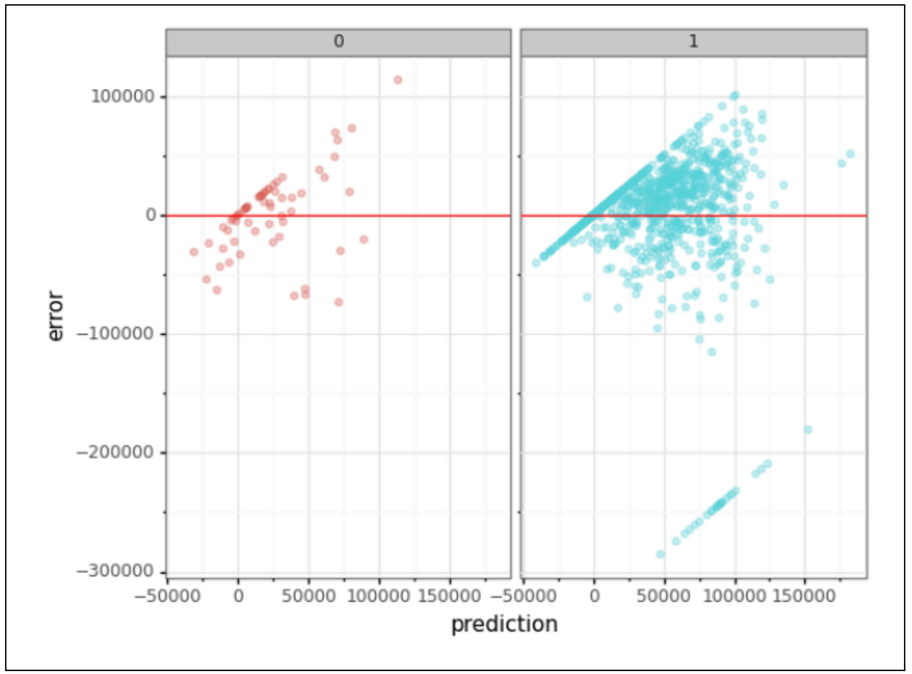




2. How well can a model predict Annual Income?

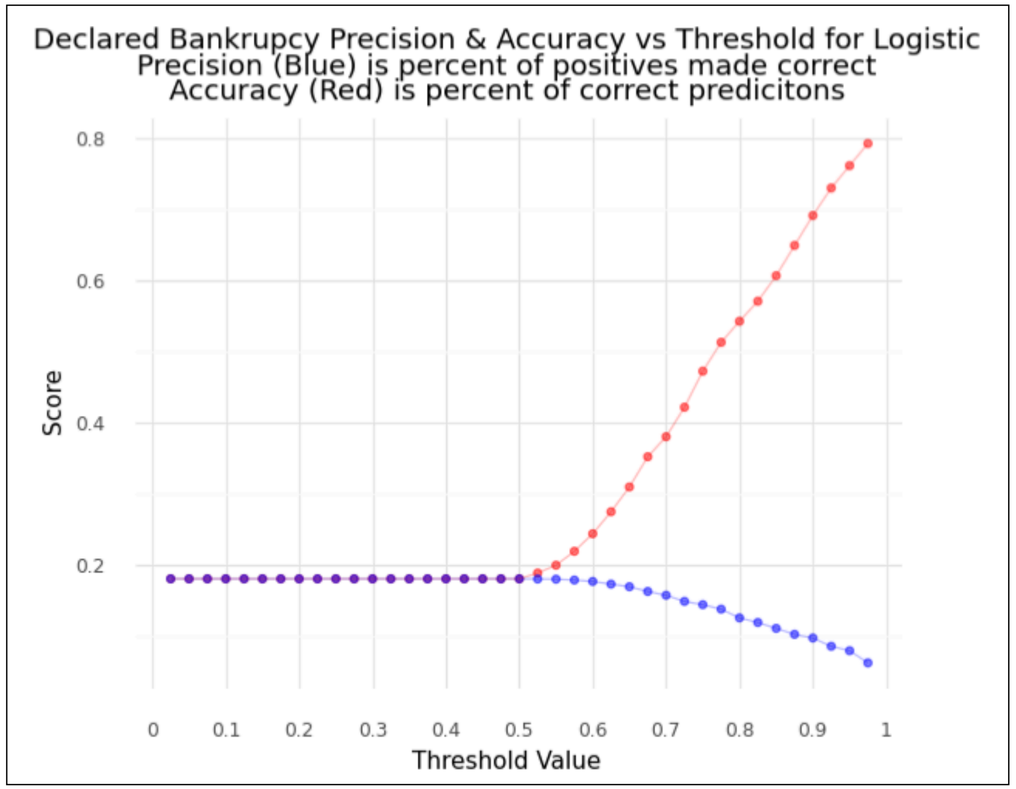
Lasso Model Summary

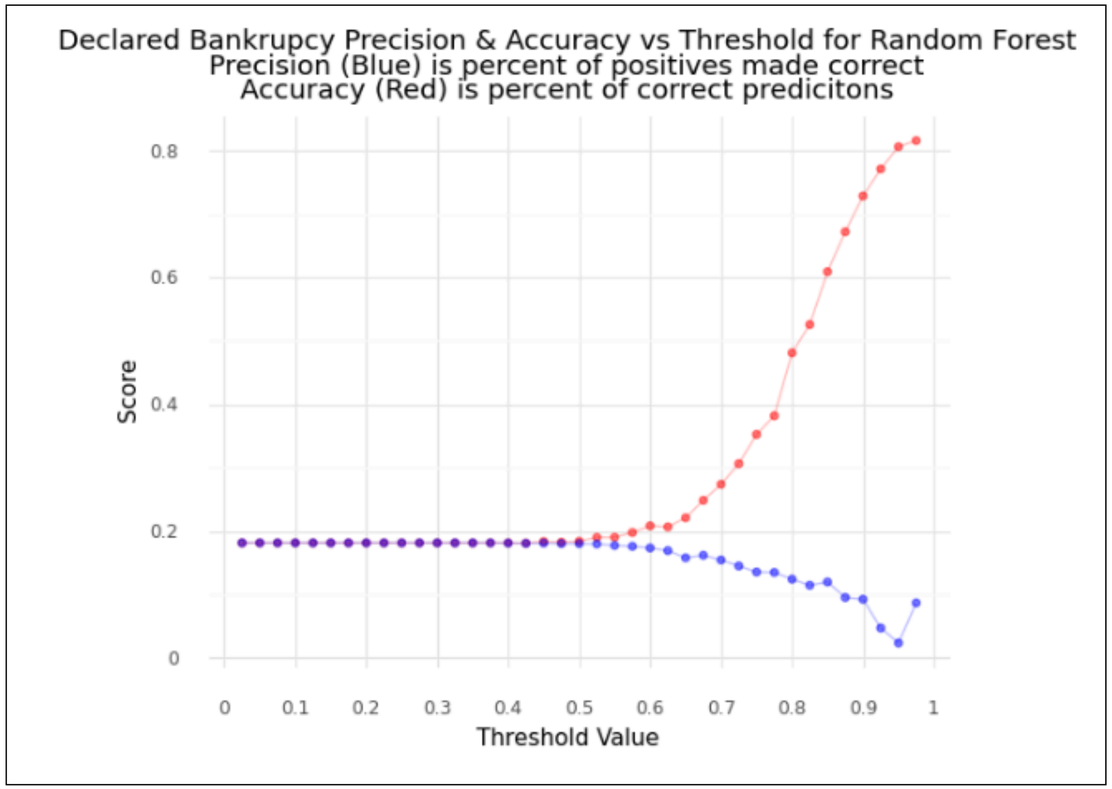




2. How well can a model predict Annual Income?

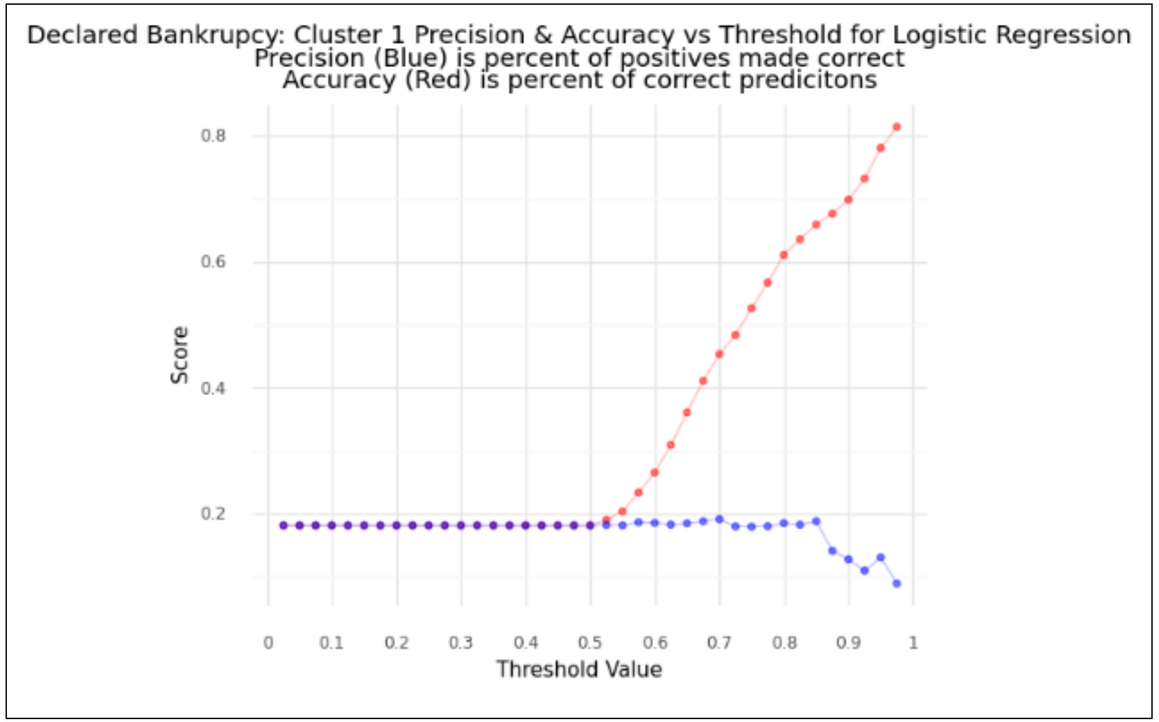
Classification Algorithm Evaluation

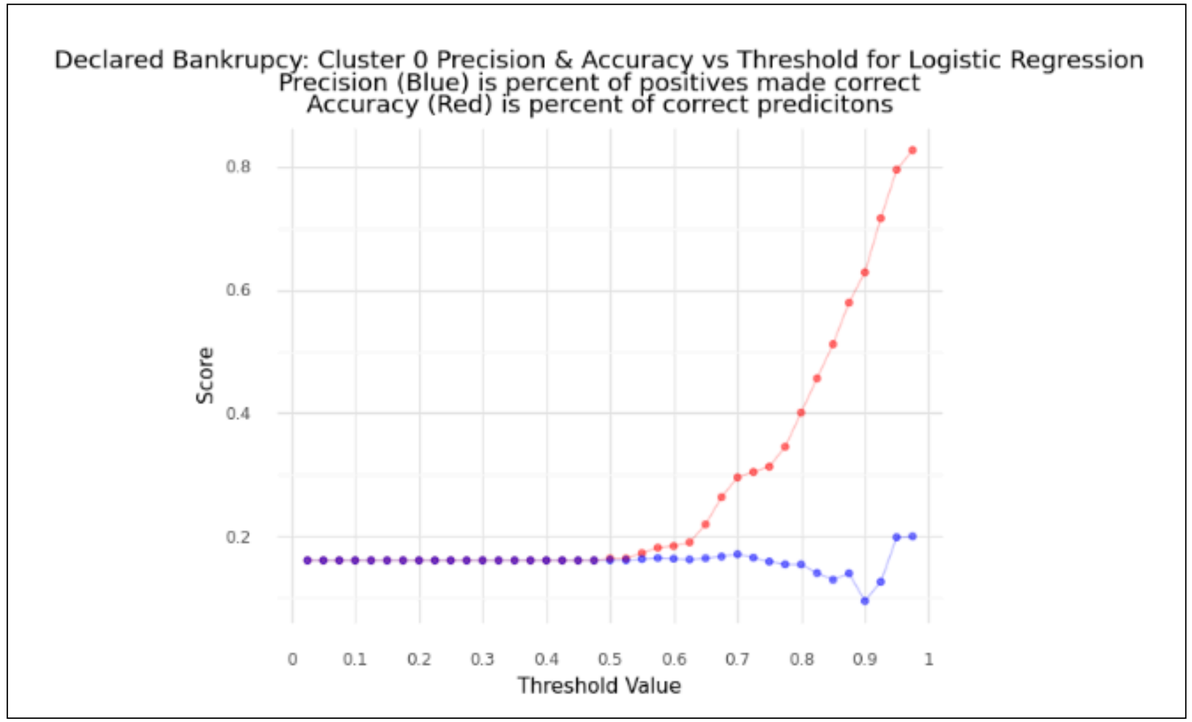




3. Can you predict who has gone bankrupt?

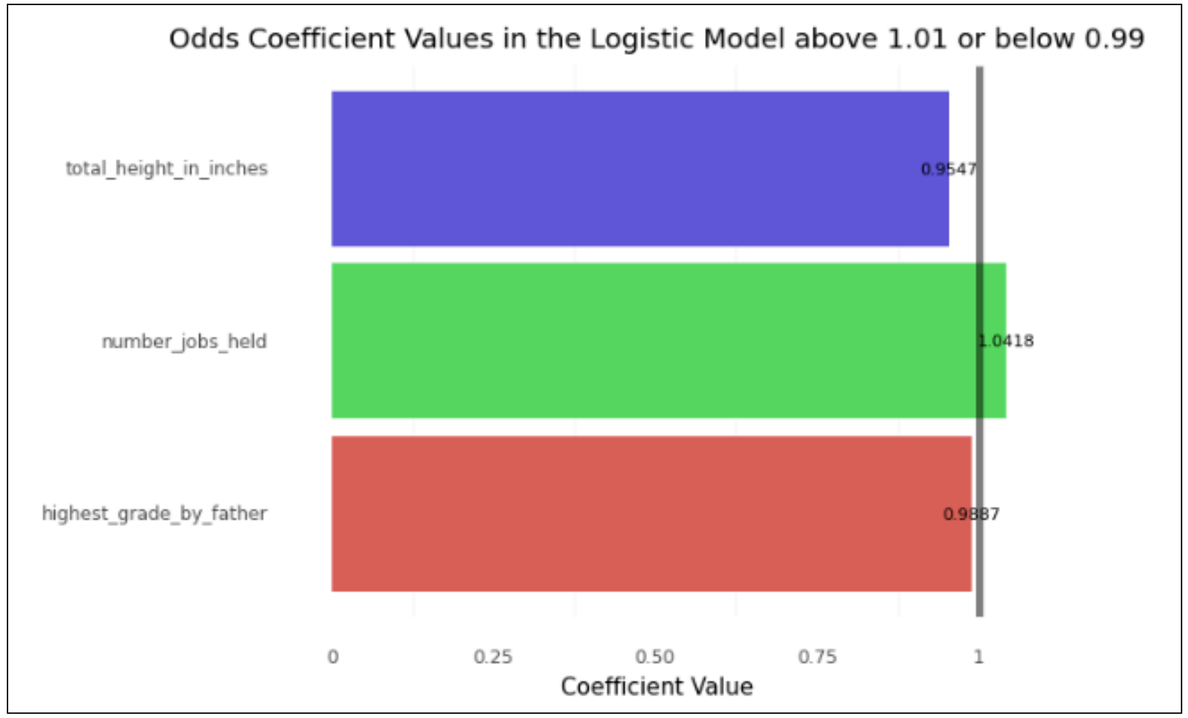
Logisitc Classification on Clusters

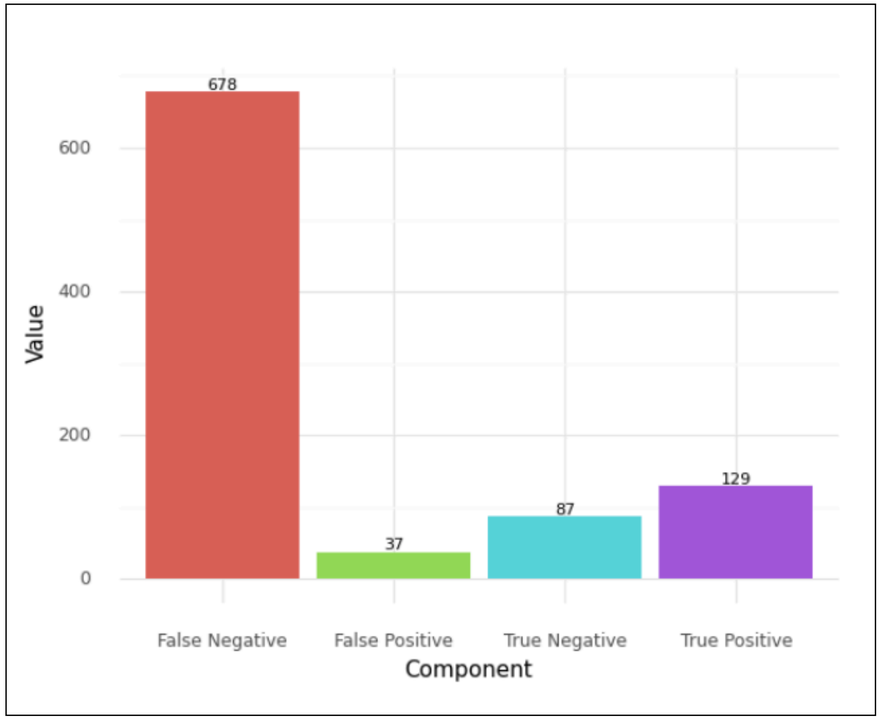




3. Can you predict who has gone bankrupt?

Logistic Regression Results





3. Can you predict who has gone bankrupt?

Summary

Moving Forward

- Using more Categorical Data
- Replacing Missing Values
- Using More Census Data for Missing Variables