PREDICTORS OF POVERTY IN THE U.S.

Insights into predictors of county poverty and child poverty.

ABSTRACT

We attempt to explore the predictors of the poverty based on the 2017 U.S. Census Data.

Anusha Sagi, Fatima Kahack, & Lydia Tse

CS 418 Final Project

Table of Contents

Premise

While poverty has long been a concern amongst philanthropists, economists, and politicians alike, the seemingly widening income gap and stagnant wages further serve as the impetus for developing modern solutions for poverty. According to a 2018 published study from the United States Census Bureau, the national poverty rate was 11.8% and 16.2% for children under age 18 (Semega, Kollar and Mohanty 2019). Poverty is inextricably linked to a variety of factors such as race, gender, or age. Thus, by utilizing data science techniques, we aim to gain better insight into the factors that can best predict whether or not a county is predicted to be categorized as falling above or below the poverty line.

Data Science Approach

Data Collection

We will use a dataset that is the compilation of 2017 U.S. Census Demographic Data (Kaggle 2017). This dataset provides ample numerical and categorical variables that can be utilized as predictors for our regression and classification models. Such predictors include race, gender, household income, unemployment rate, and so forth. Additionally, the dataset is fairly complete in the sense that missing data is at a minimum.

Data Preparation

To prepare our dataset, we performed the following.

1. Partitioned data such that 75% of the observations are utilized for training and 25% are utilized for testing.
2. Observed the following
    a. The dimensions of our training set were (2427, 37)
    b. Datatypes included float64 for 25 variables, int64 for 10 variables, and object for 2 variables
3. Determined irrelevant attributes
    a. We determined that neither "IncomeErr" now "IncomePerCapErr", income error and income per capita error respectively, were relevant to our solution. Thus, these attributes were dropped
    b. We also removed "Pacific", "Transit and FamilyWork" attribute since many of the values were 0's meaning that those attributes were not particularly meaningful
    c. "Employed" was also dropped because we already have a percent unemployed attribute which meant that this column would provide redundant information
4. Identified missing values
    a. There was only one missing value for a single observation of "Child Poverty". So, we filled this value with 0 since it was unnecessary to discard the entire observation.

Attribute	Number of Missing Values
CountyId	0

State	0
County	0
TotalPop	0
Men	0
Women	0
Hispanic	0
White	0
Black	0
Native	0
Asian	0
Pacific	0
VotingAgeCitizen	0
Income	0
IncomePerCap	0
Poverty	0
ChildPoverty	1
Professional	0
Service	0
Office	0
Construction	0
Production	0
Drive	0
Carpool	0
Transit	0
Walk	0
OtherTransp	0
WorkAtHome	0
MeanCommute	0
Employed	0
PrivateWork	0
PublicWork	0
SelfEmployed	0
FamilyWork	0
Unemployment	0

5. Added a “percent_women” attribute while dropping the count of men and women. This was done to prevent population skewing of data. Percentages also allowed us to reduce the total number of attributes in the dataset
6. Performed mean and median poverty and child poverty calculations. This allows us to better categorize whether or not a county falls above or below the poverty line. We use binary indicators for the “Poverty Category” and the “Child Poverty Category” where 0 represents falling below the mean poverty line (poverty is low) and 1 represents falling above the line (poverty is high)
7. Rearranged columns so that similar attributes are side-by-side

Data Exploration

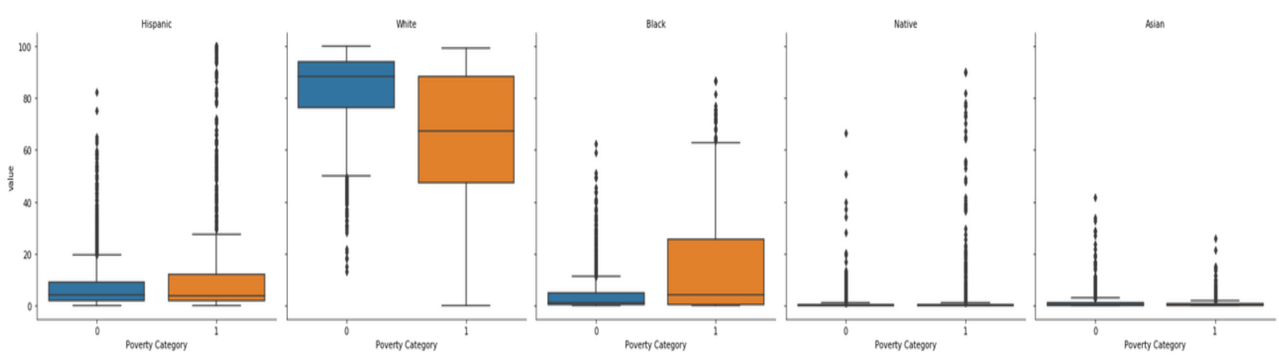
Descriptive Statistics

We begin by calculating the descriptive statistics for each attribute.

	count	mean	std	min	25%	50%	75%	max
CountyId	2427.0	31584.119489	16286.815957	1001.000000	19046.000000	30059.000000	47018.000000	7.215300e+04
TotalPop	2427.0	100277.147507	338543.745227	74.000000	11264.000000	25625.000000	66275.000000	1.010572e+07
Percent_Women	2427.0	49.997184	2.338260	19.166215	49.495212	50.432776	51.144426	5.663391e+01
Hispanic	2427.0	11.403296	19.532931	0.000000	2.100000	4.100000	9.900000	9.990000e+01
White	2427.0	74.983890	23.100311	0.000000	63.450000	83.800000	92.800000	1.000000e+02
Black	2427.0	8.547796	14.205141	0.000000	0.500000	1.900000	8.950000	8.690000e+01
Native	2427.0	1.732674	7.250648	0.000000	0.100000	0.300000	0.600000	9.030000e+01
Asian	2427.0	1.282118	2.628246	0.000000	0.200000	0.600000	1.200000	4.180000e+01
VotingAgeCitizen	2427.0	70967.882159	218772.977868	59.000000	8524.000000	19425.000000	50359.000000	6.218279e+06
Income	2427.0	49057.969922	13641.123656	11680.000000	40893.000000	47873.000000	55613.000000	1.175150e+05
IncomePerCap	2427.0	25711.307787	6629.901586	7047.000000	21669.500000	25208.000000	29038.000000	6.952900e+04
Professional	2427.0	31.536259	6.542366	11.400000	27.300000	30.500000	34.900000	6.900000e+01
Service	2427.0	18.133704	3.699940	0.000000	15.700000	17.800000	20.100000	4.640000e+01
Office	2427.0	21.897775	3.181856	4.800000	20.000000	22.100000	23.900000	3.720000e+01
Construction	2427.0	12.625010	4.156398	0.000000	9.800000	12.200000	14.800000	3.640000e+01
Production	2427.0	15.807499	5.846940	0.000000	11.450000	15.100000	19.600000	4.870000e+01
Drive	2427.0	79.589864	7.742675	4.600000	77.250000	81.000000	84.100000	9.720000e+01
Carpool	2427.0	9.842027	2.981947	0.000000	8.000000	9.500000	11.200000	2.930000e+01
Walk	2427.0	3.271240	3.941989	0.000000	1.400000	2.300000	3.900000	5.920000e+01
OtherTransp	2427.0	1.622744	1.758607	0.000000	0.850000	1.300000	1.900000	4.320000e+01
WorkAtHome	2427.0	4.736918	3.062729	0.000000	2.900000	4.100000	5.700000	2.960000e+01
MeanCommute	2427.0	23.479810	5.699332	5.100000	19.600000	23.200000	26.900000	4.510000e+01

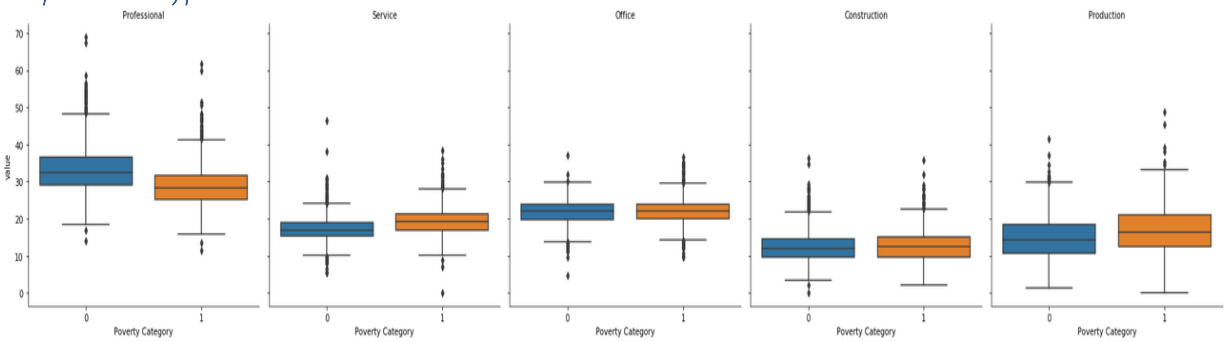
Box Plots

Race & Ethnicity Attributes



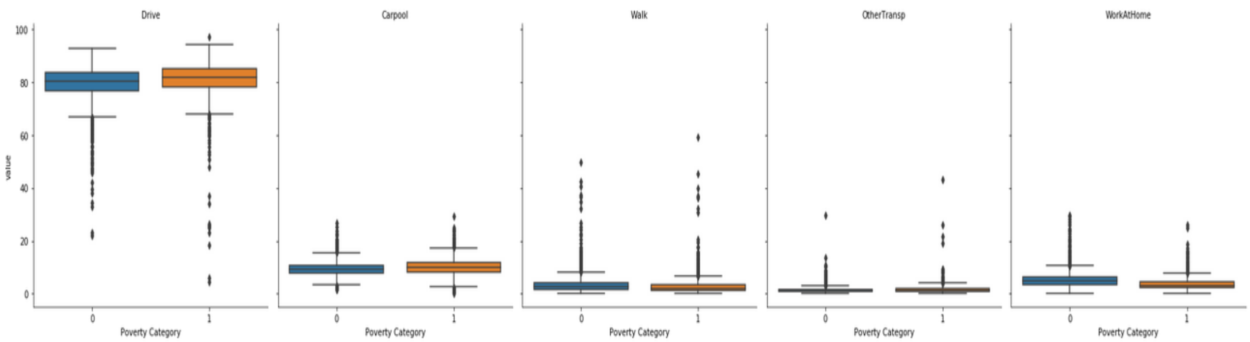
From these plots we can identify that White, Black attributes have significant differences in the poverty categories, whereas Hispanic attribute have small differences in the poverty category. There are very low percentage values for Native and Asian attributes.

Occupational Type Attributes



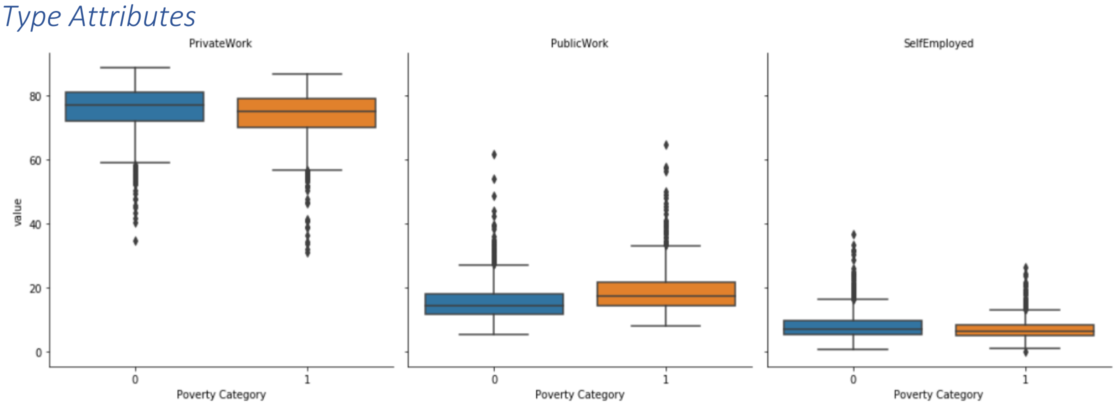
For these plots, we can identify differences in Professional, Service, and Production attributes. There are no visible differences seen in Office and Construction attributes for poverty category.

Commute Attribute



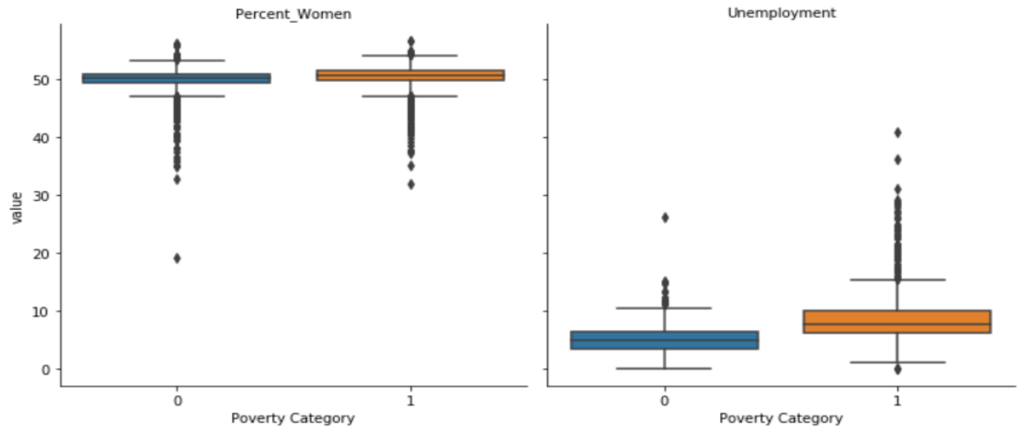
For these plots, we can cannot identify many differences in the commute attributes for poverty category.

Work Type Attributes



For these plots, there are small differences in Private work and Public Work. Likewise, there are very small difference in the Self Employed attribute of the Poverty Category.

Percent of Women in Population & Unemployment



For these plots, there is small differences in Percent_women and huge difference in unemployment attributes of Poverty Category.

Box Plot Conclusion

Through the analysis from the boxplots, we have dropped columns such as Native, Asian, Office, Construction, Drive, Walk, OtherTransport and VotingAgeCitizen attributes since they do not provide meaningful information into poverty prediction.

Data Outliers

```
Black                     341
Carpool                   108
ChildPoverty               83
Child_Poverty Category      0
County                      0
CountyId                    0
Hispanic                  319
Income                    131
IncomePerCap              112
MeanCommute                36
Percent_Women             215
Poverty                   100
Poverty Category            0
PrivateWork                72
Production                 14
Professional               76
PublicWork                 82
SelfEmployed              118
Service                    71
State                       0
TotalPop                  326
Unemployment              112
White                      97
WorkAtHome                122
dtype: int64
```

Even though this data has outliers, all of them fall within the percentage values and none of them exceed 100%. This is reasonable and hence the data provides insight in identifying and predicting the values regarding poverty.

*Note that the code regarding the aforementioned information is provided in the DPTestdata.ipynb file.

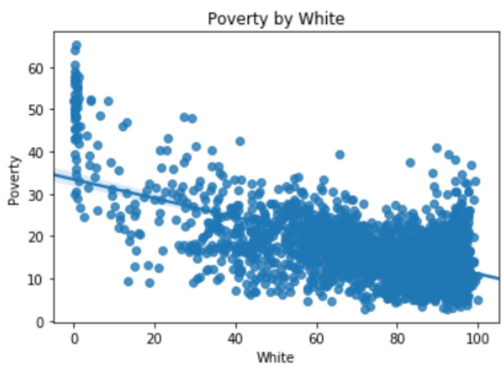
Data Modeling

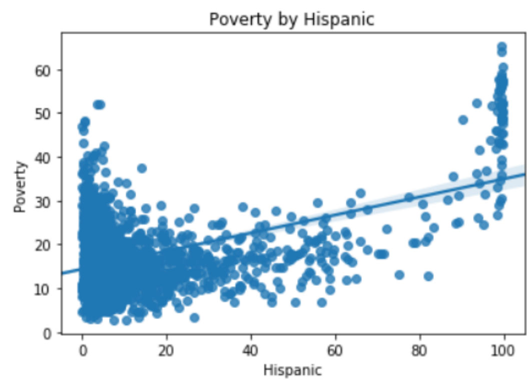
Both regression and classification models were created for this dataset.

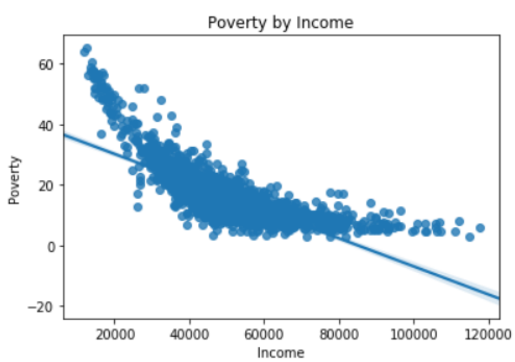
Regression Models

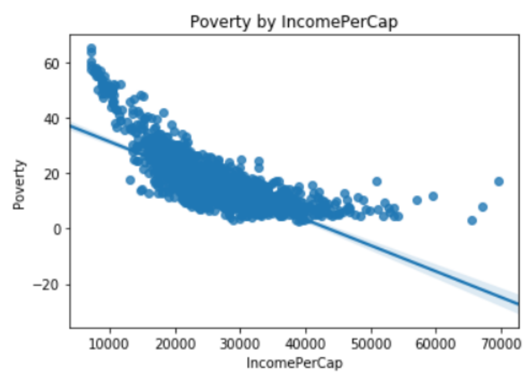
Poverty

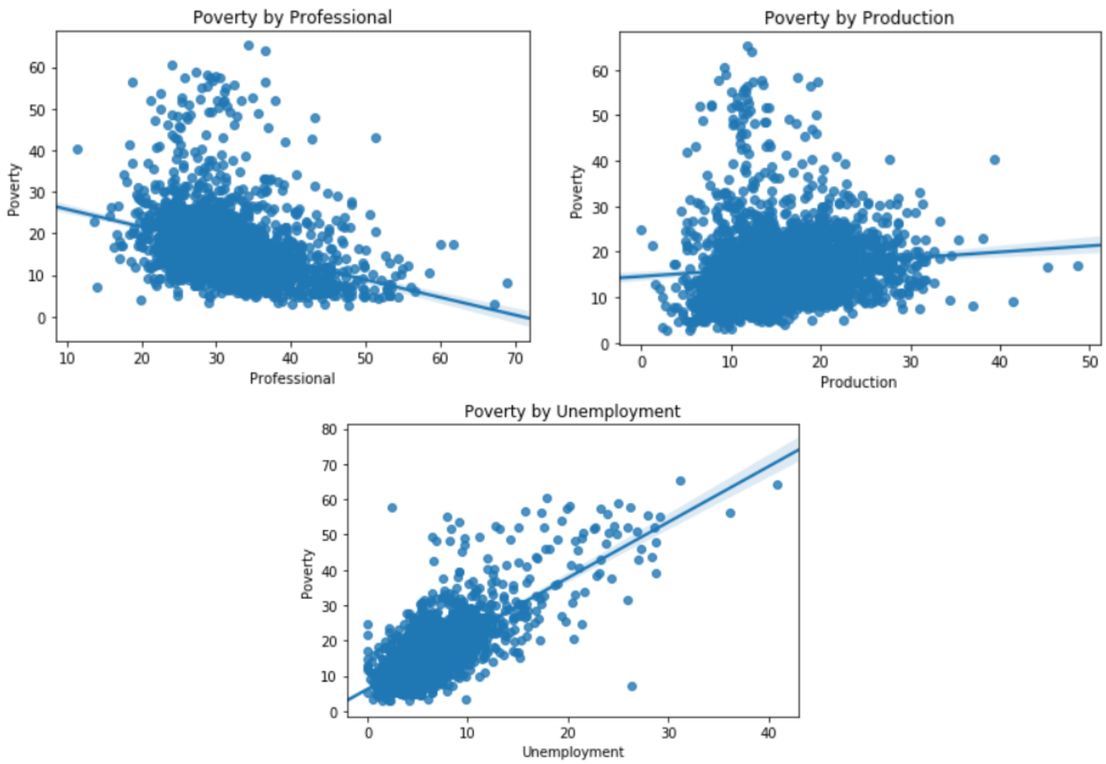
The following are plots for attributes that affect Poverty values.











Model 1

Including all variables
R squared: 0.8513298077255882
Adjusted R-squared: 0.8472980736978075
RMSE - 3.4209848294802594
Training & Validation R squared values
[0.8272983071536646, 0.8504796157247451]
The training and Validation R squared values tells us that the model is predicting validation results better than training set which is good.

Remarks: Here, using all the variables, the evaluation metrics provide a good Adjusted R squared value.

Model 2

LASSO Regression
R squared: 0.8336599709647284
Adjusted R squared: 0.829149054923094
RMSE: 3.9933497011553434

Remarks: Here using LASSO Regression, the adjusted R squared has decreased slightly but this gives us a good estimate of which parameters are more useful. In this LASSO, the attributes ‘White’, ‘Income’, ‘IncomePerCap’ and ‘Unemployment’ are considered important and affecting parameters.

Model 3

Includes: 'Percent_Women', 'White', 'Income', 'IncomePerCap', 'Professional', 'Unemployment'
R squared: 0.8471259545707231
Adjusted R squared: 0.8455972141164304
RMSE: 3.473878047933116

Remarks: Here the in combination with the LASSO attributes, we have included 'Percent_Women', 'Professional' attributes which increased the Adjusted R squared value but it is not as high as using all the variables. We want our model not to be complex, so we should find out the right set of attributes and in minimum number.

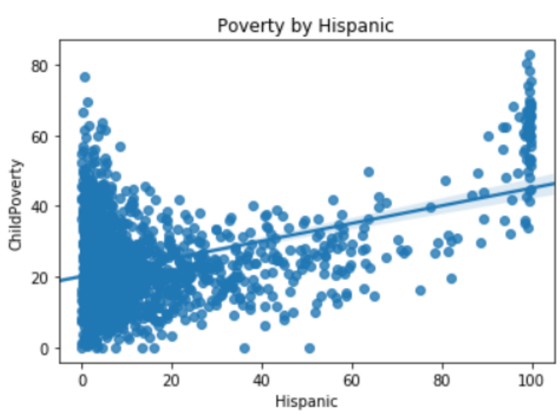
Model 4 – Best Regression Model for Poverty

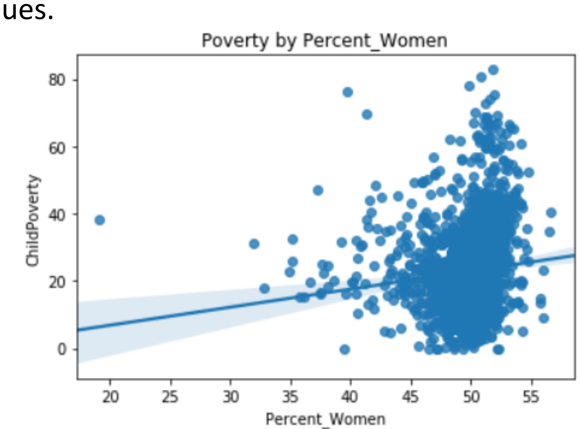
Includes: 'Hispanic', 'White', 'Income', 'IncomePerCap', 'Professional', 'Production', 'Unemployment' (BEST MODEL)
R squared: 0.8514433108569243
Adjusted R squared: 0.849707256058925
RMSE: 3.422593507053815
Training & Validation R squared values
[0.8142781696578959, 0.8503389622847697]
The model is doing very well in the validation set than the training set and has a good Adjusted R squared value.

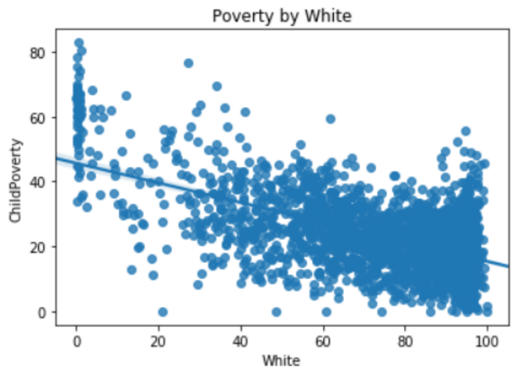
Remarks: In this model, the attributes used with LASSO and additional attributes such as 'Hispanic', 'Professional' and 'Production'

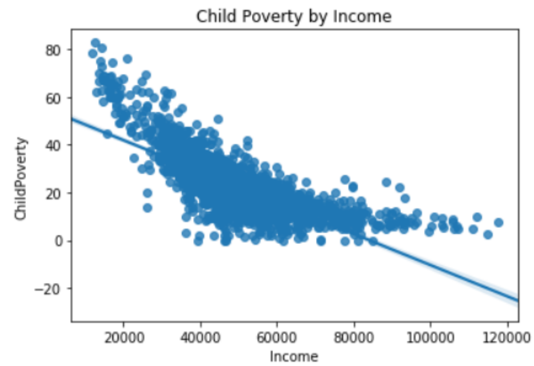
Child Poverty

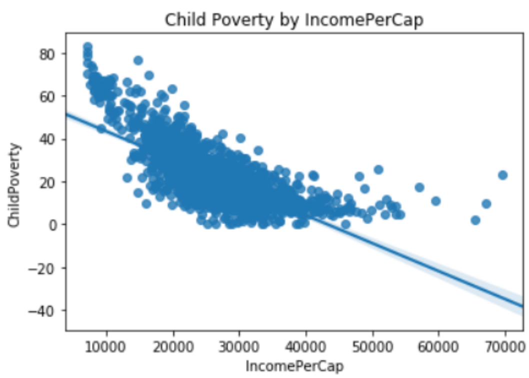
The following are plots impact Child Poverty values.

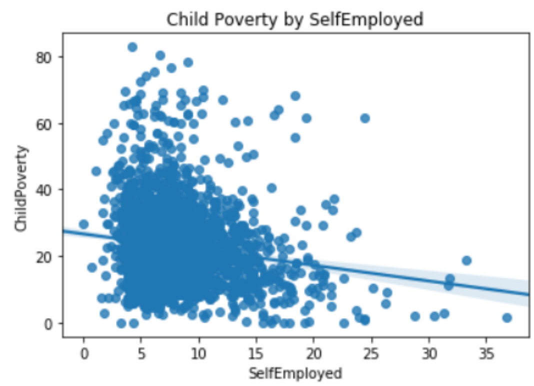


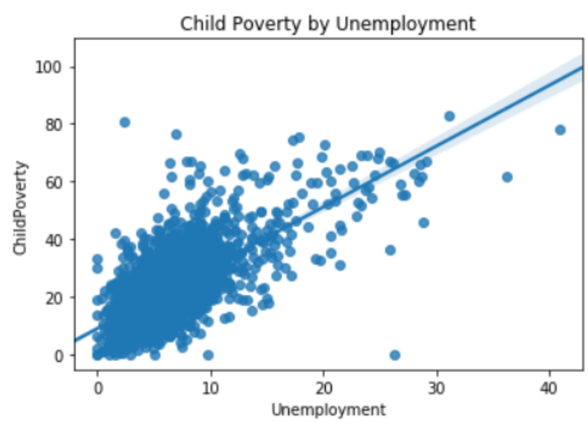












Model 1

Including all variables
R squared: 0.7725131347267581
Adjusted R squared: 0.7663439993973142
RMSE: 5.989999440186332

Remarks: Here, using all the variables, the evaluation metrics provide us a moderate Adjusted R squared value.

Model 2

LASSO Regression
R squared: 0.7680747760798697

Adjusted R squared: 0.7617852784820357
RMSE: 6.264814834157485

Remarks: In this LASSO regression model, it provides us with an important set of attributes that help in predicting the results efficiently. The attributes that were selected by the LASSO regression model are ‘White’, ‘Income’, ‘IncomePerCap’, ‘SelfEmployed’ and ‘Unemployment’.

Model 3 – Best Regression Model for Child Poverty

Includes: 'Percent_Women', 'Hispanic', 'White', 'Income', 'IncomePerCap', 'SelfEmployed', 'Unemployment' (BEST MODEL)
R squared: 0.7724443465917782
Adjusted R squared: 0.7697850985552881
RMSE: 5.992641178686748

Remarks: In this regression model, using the above-mentioned attributes in combination with LASSO attributes, we have achieved a greater Adjusted R squared value which is 0.7697. This is the best model so far for predicting Child Poverty percentage.

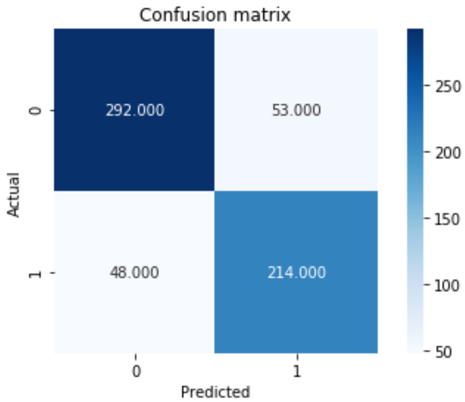
Classification Models

Poverty

In the following sections, we utilize various classification techniques to find the best model for classifying poverty.

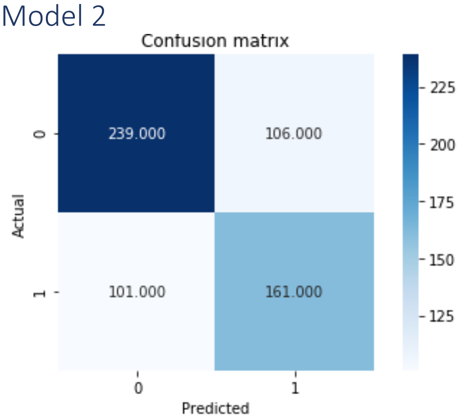
Decision Tree Classifier

Model 1



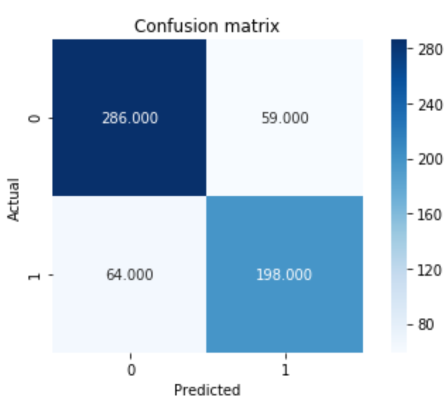
Using all variables and entropy
Number of Nodes: 295
Accuracy: 0.83
Error: 0.17
Precision: array([0.86, 0.80])
Recall: array([0.85, 0.82])
F1_Score: array([0.85, 0.81])

Model 2



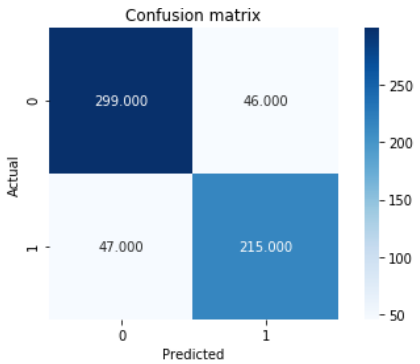
Using some variables and entropy - Hispanic, White, Black, Percent_Women
# of Nodes: 879
Accuracy: 0.66
Error: 0.34
Precision: array([0.70, 0.60])
Recall: array([0.69, 0.61])
F1_Score: array([0.70, 0.61])

Model 3



Using some variables and entropy – Unemployment, Income, SelfEmployed
# of Nodes: 577
Accuracy: 0.80
Error: 0.20
Precision: array([0.82, 0.77])
Recall: array([0.83, 0.76])
F1_Score: array([0.82, 0.76])

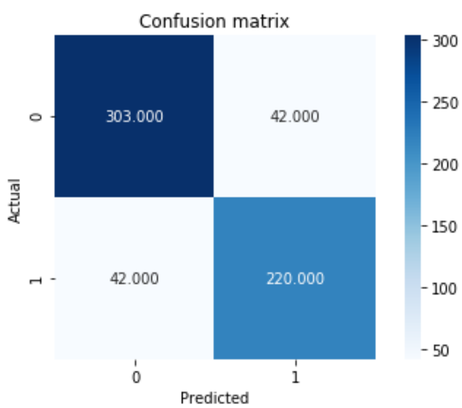
Model 4



Using some variables and entropy - Hispanic, Black, White, Percent_Women, Unemployment, Income, SelfEmployed

# of Nodes: 405
Accuracy: 0.85
Error: 0.15
Precision: array([0.86, 0.82])
Recall: array([0.87, 0.82])
F1_Score: array([0.87, 0.82])

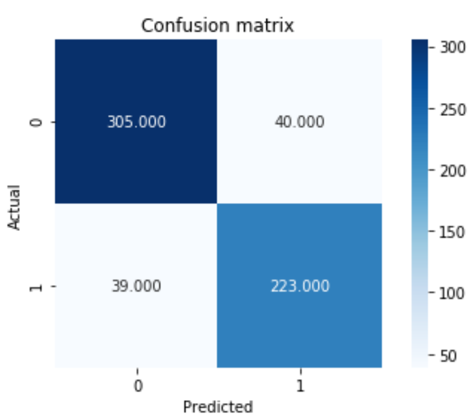
Model 5



Using some variables and entropy - Hispanic, Black, White, Percent_Women, Unemployment, Income, SelfEmployed, Professional

# of Nodes: 375
Accuracy: 0.86
Error: 0.14
Precision: array([0.88, 0.84])
Recall: array([0.88, 0.84])
F1_Score: array([0.88, 0.84])

Model 6

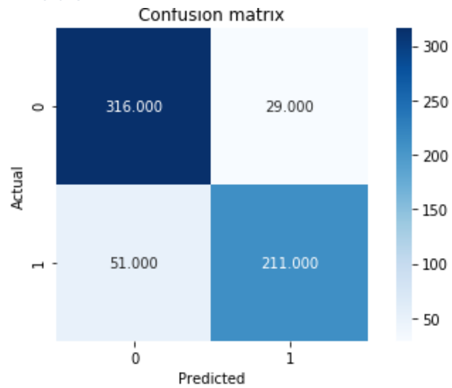


Using some variables and entropy - Hispanic, Black, White, Unemployment, Income, SelfEmployed, Professional

# of Nodes: 381
Accuracy: 0.87
Error: 0.13
Precision: array([0.89, 0.85])
Recall: array([0.88, 0.85])
F1_Score: array([0.89, 0.85])

K-Nearest Neighbors

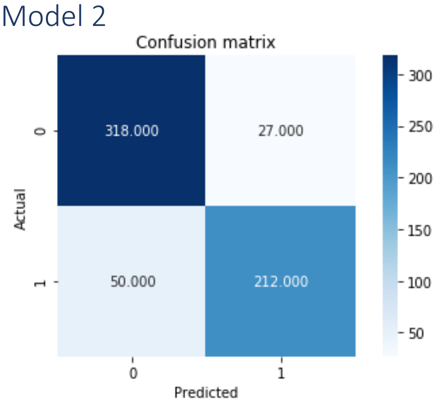
Model 1



Using some variables with k = 5 - Hispanic, Black, White, Unemployment, Income, SelfEmployed, Professional

Accuracy: 0.87
Error: 0.13
Precision: array([0.86, 0.88])
Recall: array([0.92, 0.81])
F1_Score: array([0.89, 0.84])

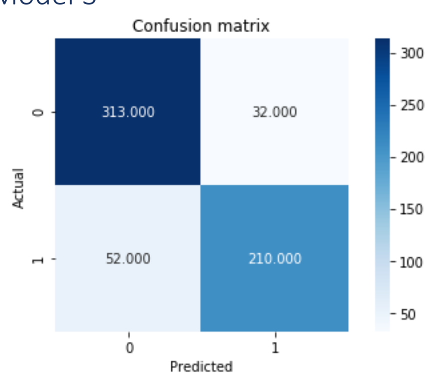
Model 2



Using some variables with k = 7 - Hispanic, Black, White, Unemployment, Income, SelfEmployed, Professional

Accuracy: 0.87
Error: 0.13
Precision: array([0.86, 0.89])
Recall: array([0.92, 0.81])
F1_Score: array([0.89, 0.85])

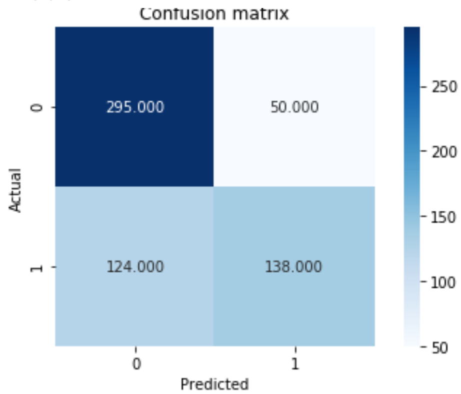
Model 3



Using some variables with k = 5 - Hispanic, Black, White, Unemployment, Income, SelfEmployed, Professional, Percent_Women

Accuracy: 0.86
Error: 0.14
Precision: array([0.86, 0.87])
Recall: array([0.91, 0.80])
F1_Score: array([0.88, 0.83])

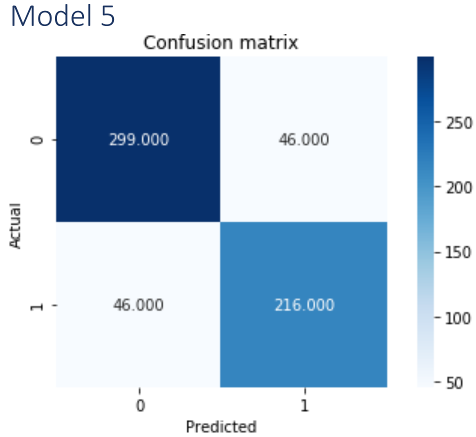
Model 4



Using some variables with k = 7 - Hispanic, Black, White, Percent_Women

Accuracy: 0.71
Error: 0.29
Precision: array([0.70, 0.73])
Recall: array([0.86, 0.53])
F1_Score: array([0.77, 0.61])

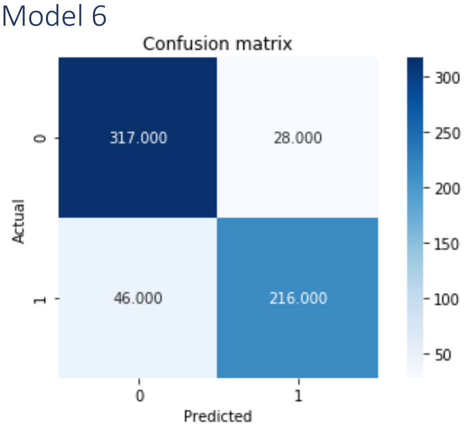
Model 5



Using some variables with k = 7 - Unemployment, Income, SelfEmployed, Professional

Accuracy: 0.85
Error: 0.15
Precision: array([0.87, 0.82])
Recall: array([0.87, 0.82])
F1_Score: array([0.87, 0.82])

Model 6

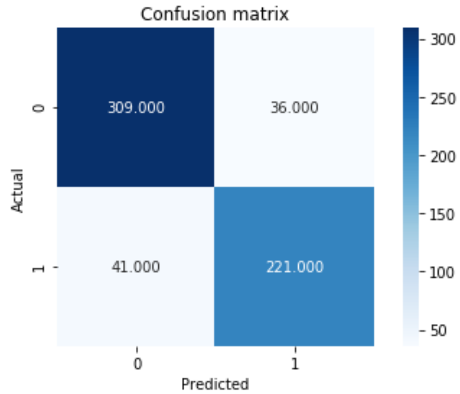


Using some variables with k = 7 - Unemployment, Income, SelfEmployed, Professional, Percent_Women

Accuracy: 0.88
Error: 0.13
Precision: array([0.87, 0.88])
Recall: array([0.92, 0.82])
F1_Score: array([0.90, 0.85])

Support Vector Machines

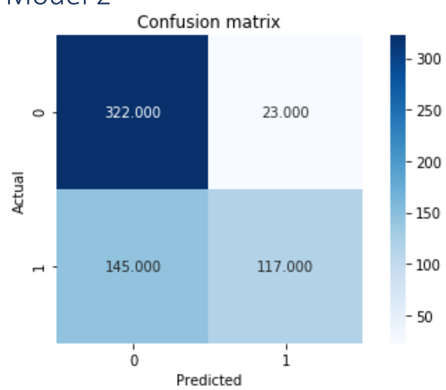
Model 1



Using all variables

Accuracy: 0.87
Error: 0.13
Precision: array([0.88, 0.86])
Recall: array([0.90, 0.84])
F1_Score: array([0.89, 0.85])

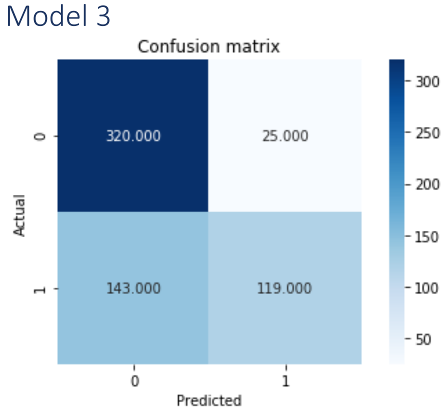
Model 2



Using some variables - Hispanic, White, Black

Accuracy: 0.72
Error: 0.28
Precision: array([0.69, 0.84])
Recall: array([0.93, 0.45])
F1_Score: array([0.79, 0.58])

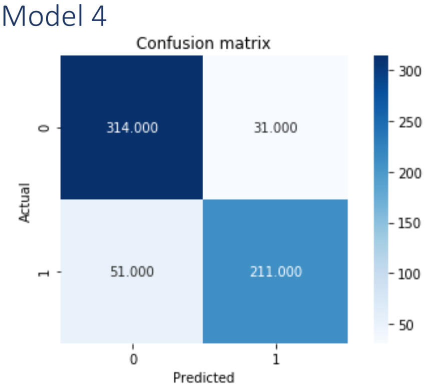
Model 3



Using some variables - Hispanic, White, Black, Percent_Women

Accuracy: 0.72
Error: 0.28
Precision: array([0.69, 0.83])
Recall: array([0.93, 0.45])
F1_Score: array([0.79, 0.59])

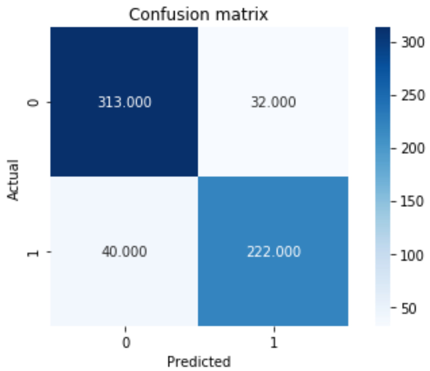
Model 4



Using some variables - Unemployment, Income, SelfEmployed

Accuracy: 0.86
Error: 0.14
Precision: array([0.86, 0.87])
Recall: array([0.91, 0.80])
F1_Score: array([0.88, 0.84])

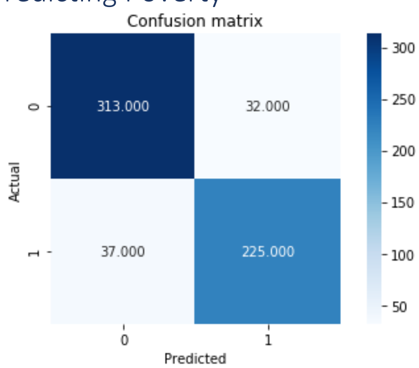
Model 5



Using some variables - Unemployment, Income, SelfEmployed, Black, White, Hispanic, Percent_Women

Accuracy: 0.88
Error: 0.12
Precision: array([0.88, 0.87])
Recall: array([0.91, 0.85])
F1_Score: array([0.90, 0.86])

Model 6 – Best Classification Model for Predicting Poverty



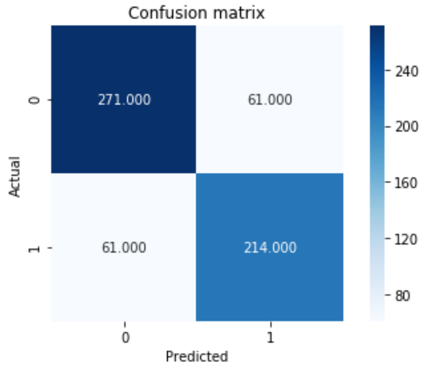
Using some variables - Unemployment, Income, SelfEmployed, Black, White, Hispanic, Percent_Women, Professional

Accuracy: 0.89
Error: 0.11
Precision: array([0.89, 0.88])
Recall: array([0.91, 0.86])
F1_Score: array([0.90, 0.87])

Child Poverty

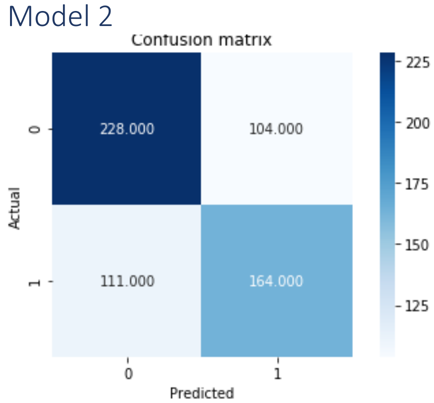
Decision Tree Classifier

Model 1



Using all variables and entropy
# of Nodes: 365
Accuracy: 0.80
Error: 0.20
Precision: array([0.82, 0.78])
Recall: array([0.82, 0.78])
F1_Score: array([0.82, 0.78])

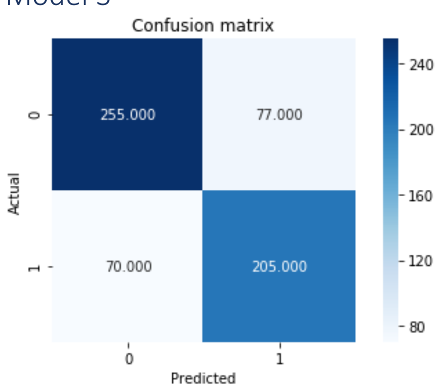
Model 2



Using some variables and entropy - Hispanic, White, Black, Percent_Women
# of Nodes: 887

Accuracy: 0.65
Error: 0.35
Precision: array([0.67, 0.61])
Recall: array([0.69, 0.60])
F1_Score: array([0.68, 0.60])

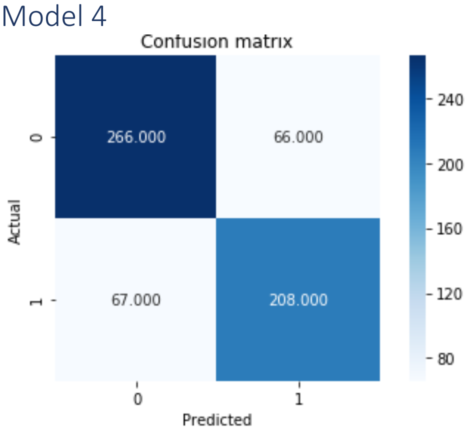
Model 3



Using some variables and entropy - Unemployment, Income, SelfEmployed
# of Nodes: 673

Accuracy: 0.76
Error: 0.24
Precision: array([0.78, 0.73])
Recall: array([0.77, 0.75])
F1_Score: array([0.78, 0.74])

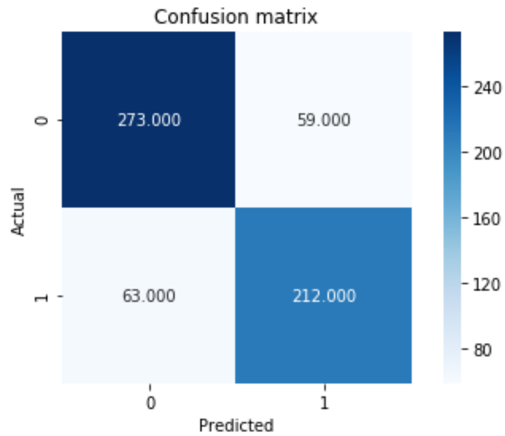
Model 4



Using some variables and entropy - Hispanic, Black, White, Percent_Women, Unemployment, Income, SelfEmployed
# of Nodes: 435

Accuracy: 0.78
Error: 0.22
Precision: array([0.80, 0.76])
Recall: array([0.80, 0.76])
F1_Score: array([0.80, 0.76])

Model 5

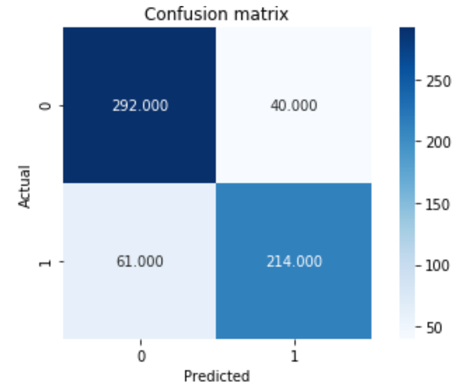


Using some variables and entropy - Hispanic, Black, White, Unemployment, Income, SelfEmployed
# of Nodes: 411

Accuracy: 0.80
Error: 0.20
Precision: array([0.81, 0.78])
Recall: array([0.82, 0.77])
F1_Score: array([0.82, 0.78])

K-Nearest Neighbors

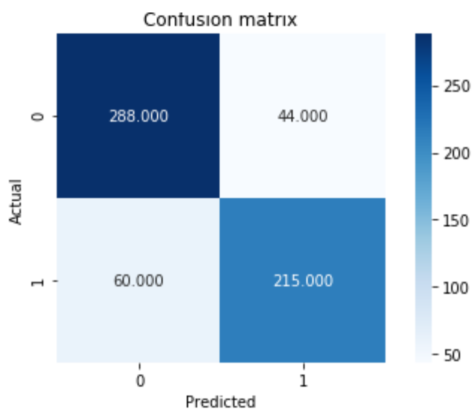
Model 1



Using all variables with k = 13

Accuracy: 0.83
Error: 0.17
Precision: array([0.83, 0.84])
Recall: array([0.88, 0.78])
F1_Score: array([0.85, 0.81])

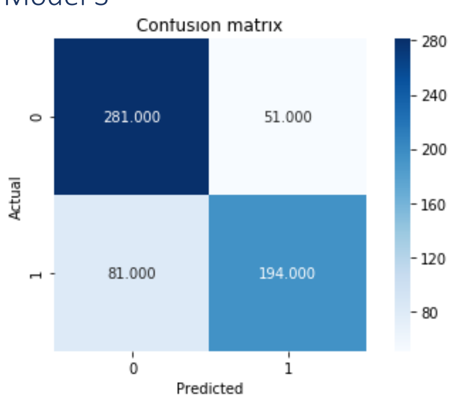
Model 2



Using some variables with k = 13 - Hispanic, Black, White, Unemployment, Income, SelfEmployed, Professional

Accuracy: 0.83
Error: 0.17
Precision: array([0.83, 0.83])
Recall: array([0.87, 0.78])
F1_Score: array([0.85, 0.81])

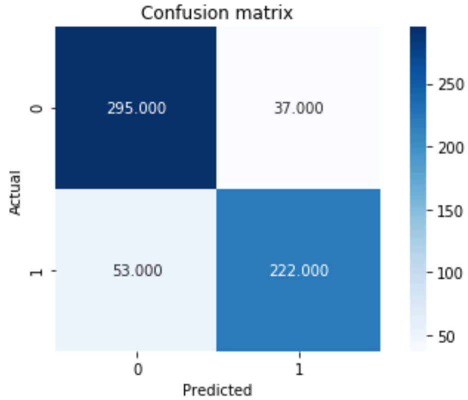
Model 3



Using some variables with k = 13 - Hispanic, Black, White, Percent_Women

Accuracy: 0.78
Error: 0.22
Precision: array([0.78, 0.79])
Recall: array([0.85, 0.71])
F1_Score: array([0.81, 0.75])

Model 4 – Best Model Classification model for predicting Child Poverty

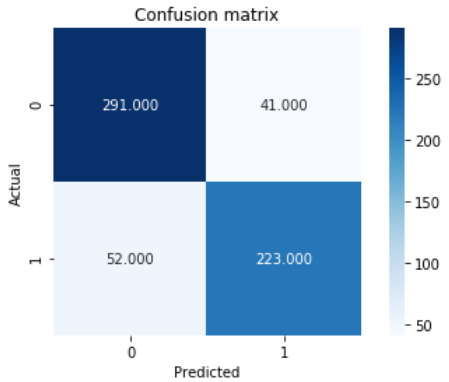


Using some variables with k = 13 - Hispanic, Black, White, Unemployment, Income, WorkAtHome

Accuracy: 0.85
Error: 0.15
Precision: array([0.85, 0.86])
Recall: array([0.89, 0.81])
F1_Score: array([0.87, 0.83])

Support Vector Machines

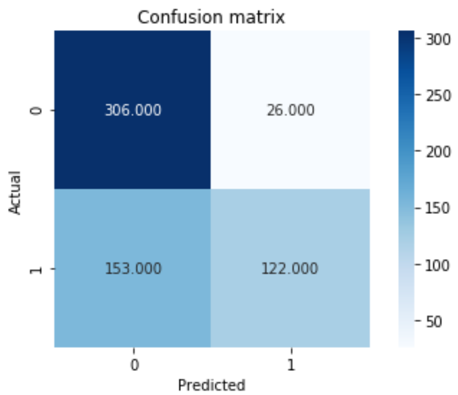
Model 1



Using all variables

Accuracy: 0.85
Error: 0.15
Precision: array([0.85, 0.84])
Recall: array([0.88, 0.81])
F1_Score: array([0.86, 0.83])

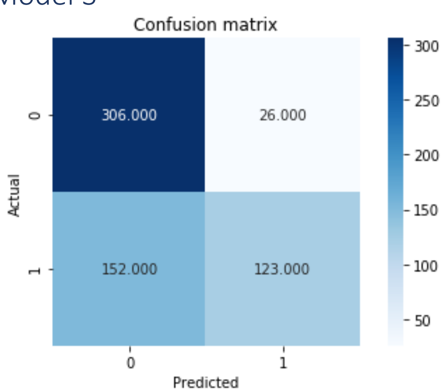
Model 2



Using some variables- Hispanic, White, Black

Accuracy: 0.71
Error: 0.29
Precision: array([0.67, 0.82])
Recall: array([0.92, 0.44])
F1_Score: array([0.77, 0.58])

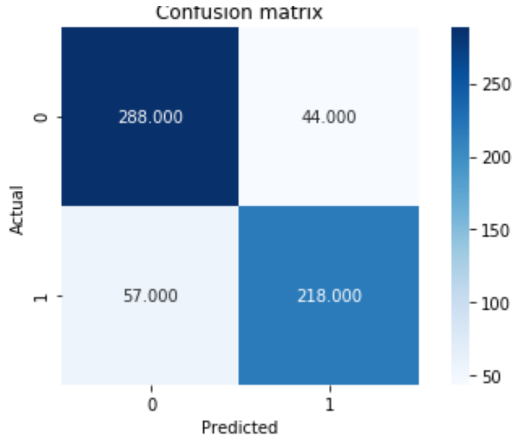
Model 3



Using some variables- Hispanic, White, Black, Percent_Women

Accuracy: 0.71
Error: 0.29
Precision: array([0.67, 0.83])
Recall: array([0.92, 0.45])
F1_Score: array([0.77, 0.58])

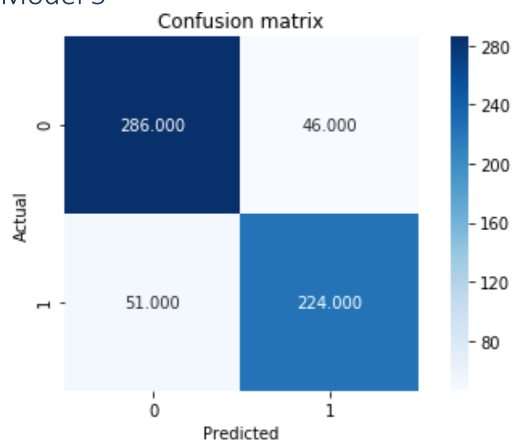
Model 4



Using some variables- Unemployment, Income, SelfEmployed

Accuracy: 0.83
Error: 0.17
Precision: array([0.83, 0.83])
Recall: array([0.87, 0.79])
F1_Score: array([0.85, 0.81])

Model 5



Using some variables- Hispanic, White, Black, Unemployment

Accuracy: 0.84
Error: 0.16
Precision: array([0.85, 0.83])
Recall: array([0.86, 0.81])
F1_Score: array([0.86, 0.82])

Conclusion

The best model that accurately classifies poverty in an area is the support vector machine model 6. This model uses variables Hispanic, Black, White, Unemployment, Income, SelfEmployed, Percent_Women, and Professional.

The best model that accurately classifies child poverty in an area is the K-Nearest Neighbors model 4. The variables used in this model are Hispanic, Black, White, Unemployment, Income, and WorkAtHome.

The variables shared among these models are Hispanic, Black, White, Unemployment, and Income.

Sample Output

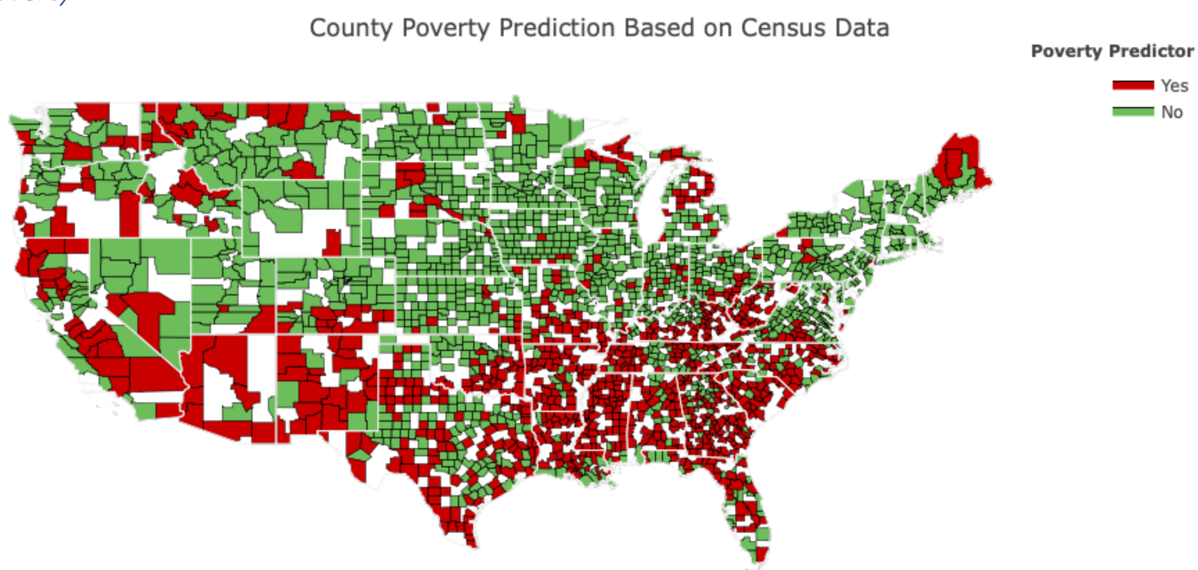
[148]:

	CountyId	State	County	Poverty	ChildPoverty	Poverty Category	Child_Poverty Category
0	1005	Alabama	Barbour County	27.773252	38.212651	1	1
1	1025	Alabama	Clarke County	27.235754	40.508661	1	1
2	1037	Alabama	Coosa County	22.358539	34.926116	1	1
3	1047	Alabama	Dallas County	31.408525	45.562577	1	1
4	1053	Alabama	Escambia County	26.052590	36.944246	1	1

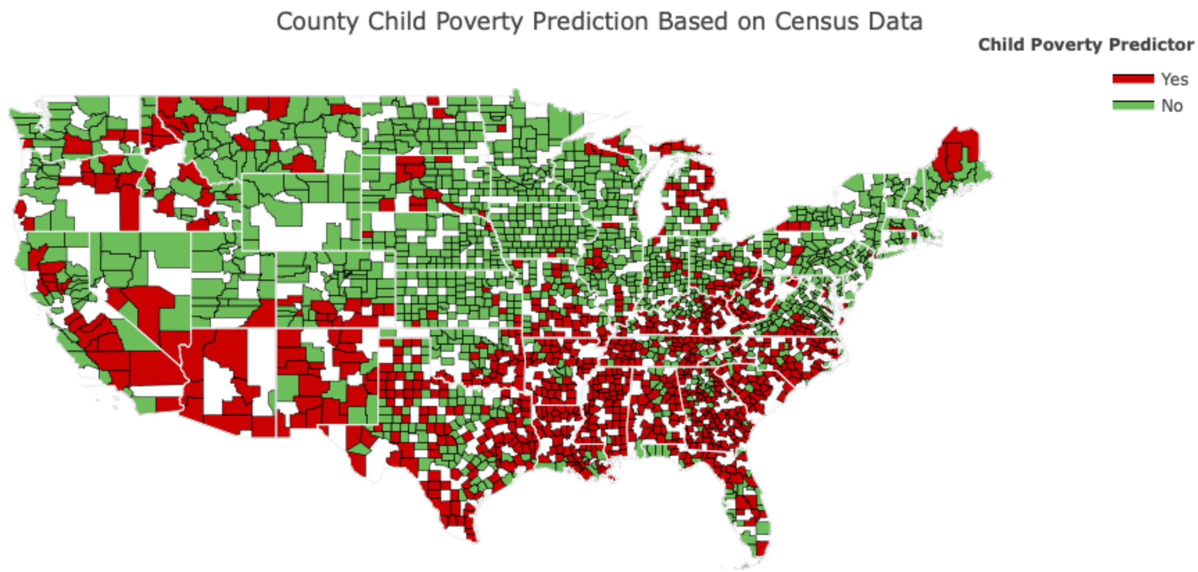
Data Visualization

Choropleth Maps

Poverty



Child Poverty



Concluding Remarks

At glance, the two maps appear identical. However, there are a few counties in which the child poverty and the poverty indicator do not coincide. In spite of this, we observe that poverty and child poverty are closely related. Finally, it is important to note that while the strongest predictors of poverty in a county involve race, we want to emphasize that this does not imply that race is a causation of poverty. Instead, we conclude that poverty is closely related to opportunities and access to employment based on the common predictors of employment and type of employment.

Bibliography

2017. *Kaggle.* 12 13. Accessed 11 22, 2019. https://www.kaggle.com/muonneutrino/us-census-demographic-data.

Semega, Jessica, Melissa Kollar, and Abinash Mohanty. 2019. "Income and Poverty in the United States: 2018." *Income and Poverty in the United States: 2018.* September 10. Accessed 11 28, 2019. https://www.census.gov/library/publications/2019/demo/p60-266.html.