Prediction of High Poverty-Stricken Schools

Derek Nguyen, Yessica Gaona, Ian Roquebert, James Helgren

TEXAS STATE UNIVERSITY

MEMBER THE TEXAS STATE UNIVERSITY SYSTEM

Introduction

- We are interested in answering if the number of students that qualify for Free/Reduced school lunches is a better indicator than Title I designation for predicting high poverty in schools.
- We gathered data from the Common Core of Data (CDD) which is the Department of Education's primary database on public elementary and secondary education in the US.
- Focus on South region of the U.S:

  Alabama, Arkansas, Delaware, Florida, Georgia, Kentucky, Luisiana, Maryland, Mississipi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia.

Questions

- We are interested in answering if the number of students that qualify for Free/Reduced school lunches is a better indicator than Title I designation for predicting high poverty in schools.
- What states have the highest percentage of strict poverty schools?
- What features are the best indication of poverty?
- How is the distribution of poverty with respect to state?

Brief Overview of the Data

- Table generator made gathering data more efficient than working with raw data files, however not as many features were available.
- Title 1 designation was not available for school years before 1998.
    - "Indicators of charter, magnet, Title I, and schoolwide Title I schools were added to CCD in 1998-99, and they are presented without further editing or imputation in the Longitudinal Database"
- We gathered data year by year then concatenated years for each section:

Ground Truth Labeling	Feature Selection	Modeling	Model Selection	Testing
1998-2003	2003-2006	2006-2015	2015-2018	2018-2020

Data Cleaning

- 19 columns before data cleaning and normalization
- Removed rows with missing values; non-numeric and not applicable data represented by:

- † indicates that the data are not applicable.
- – indicates that the data are missing.
- ‡ indicates that the data do not meet NCES data quality standards.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 88158 entries, 0 to 31928
Data columns (total 19 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   School                            88158 non-null  object
 1   State                             88158 non-null  object
 2   State Abbr                        88158 non-null  object
 3   School ID (NCES)                  88158 non-null  object
 4   Agency ID (NCES)                  88158 non-null  object
 5   School-wide Title I               88158 non-null  int64
 6   Total Students                    88158 non-null  float64
 7   Free and Reduced Lunch Students   88158 non-null  float64
 8   Male                              88158 non-null  float64
 9   Female                            88158 non-null  float64
 10  American Indian/Alaska Native     88158 non-null  float64
 11  Asian or Asian/Pacific Islander   88158 non-null  float64
 12  Black or African American         88158 non-null  float64
 13  Hispanic                          88158 non-null  float64
 14  White                             88158 non-null  float64
 15  FTE Teachers                      88158 non-null  float64
 16  Pupil/Teacher Ratio               88158 non-null  float64
 17  Year                              88158 non-null  int64
 18  Poverty Level                     88158 non-null  int64
dtypes: float64(11), int64(3), object(5)
memory usage: 13.5+ MB
None
```

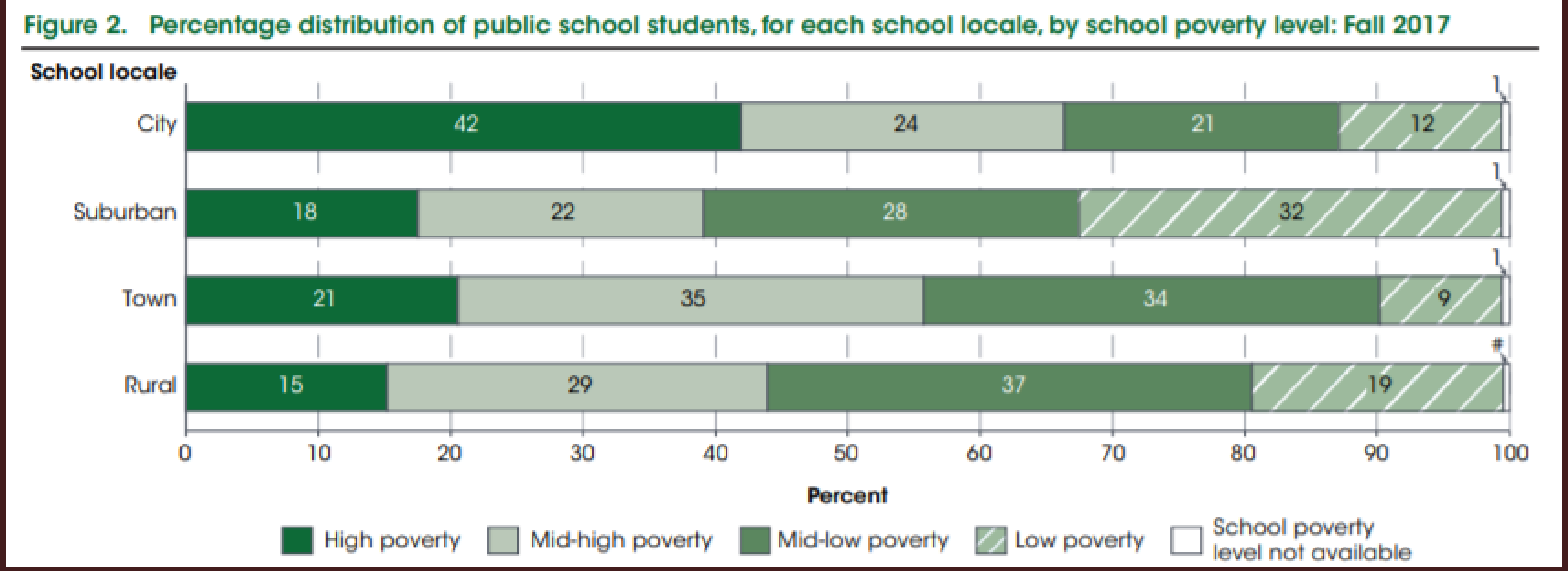
Normalization

- Normalized Features
  - Free and Reduced Lunch Students
  - Male
  - Female
  - American Indian/Alaska Native
  - Asian or Asian/Pacific Islander
  - Black or African American
  - Hispanic
  - White
- To normalize these features, we divided by 'Total Students' column, then we deleted it.
- Eliminated 'FTE Teachers' because we will use 'Pupil/Teacher Ratio' instead.

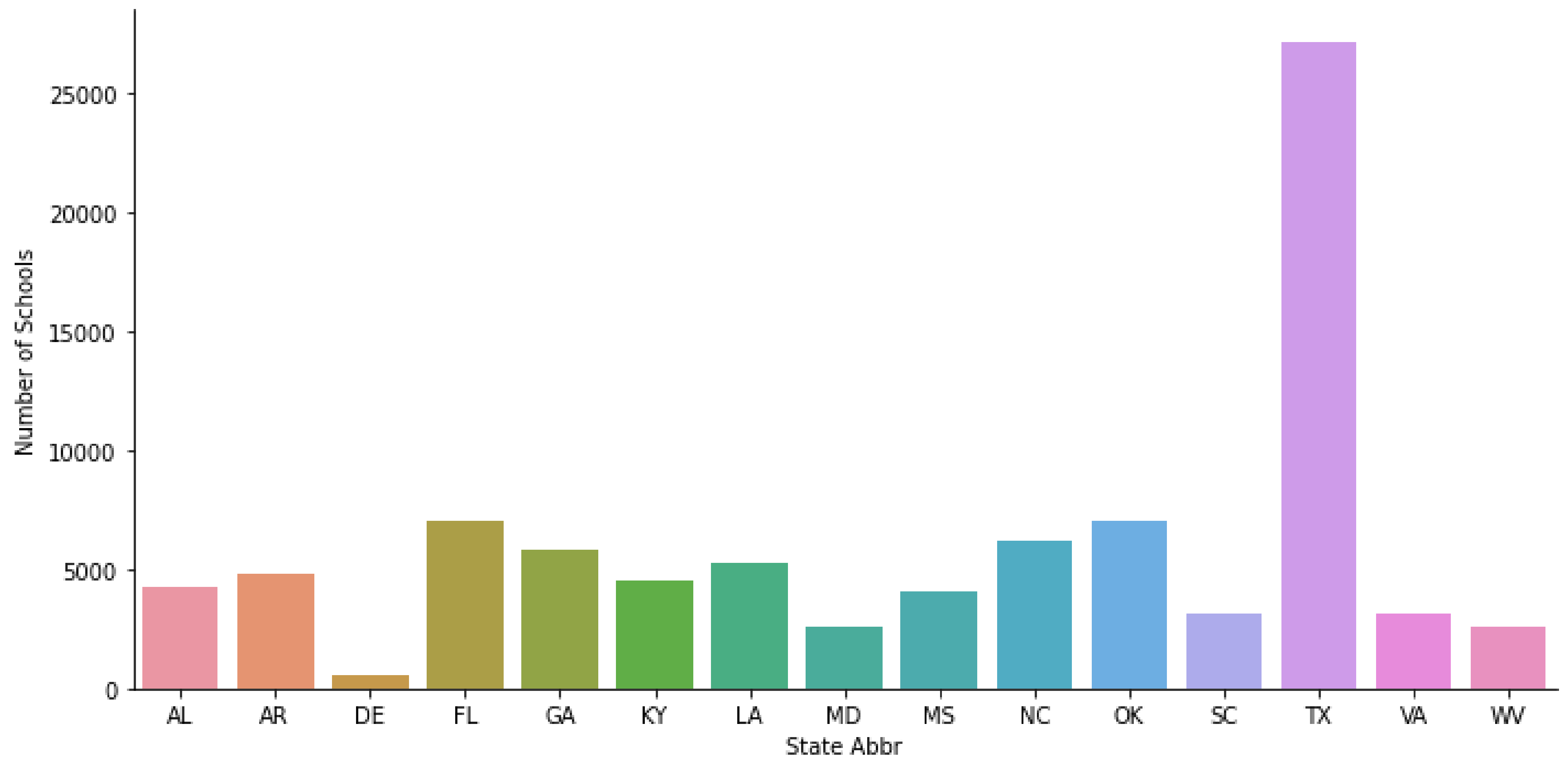
```
                            School     State State Abbr School ID (NCES)  \
0                 6TH GRADE CENTER     Texas         TX     482172005738
1        7TH AND 8TH GRADE ACADEMY  Oklahoma         OK     402097000599
2  A B CHANDLER ELEMENTARY SCHOOL  Kentucky         KY     210271000573
3                     A B DUNCAN EL     Texas         TX     481944001801
4                      A B MCBAY EL     Texas         TX     483042003424

  Agency ID (NCES)  School-wide Title I  Total Students  \
0          4821720                    1           792.0
1          4020970                    0           818.0
2          2102710                    1           285.0
3          4819440                    1           433.0
4          4830420                    1           656.0

   Free and Reduced Lunch Students      Male    Female  ...  \
0                        0.527778  0.489899  0.510101  ...
1                        0.537897  0.491443  0.508557  ...
2                        0.533333  0.491228  0.480702  ...
3                        0.748268  0.515012  0.484988  ...
4                        0.696646  0.518293  0.481707  ...

   Black or African American  Hispanic     White  FTE Teachers  \
0                   0.247475  0.146465  0.597222          49.0
1                   0.146699  0.014670  0.496333          46.5
2                   0.028070  0.000000  0.943860          15.4
3                   0.043880  0.676674  0.279446          25.6
4                   0.382622  0.169207  0.442073          39.6

   Pupil/Teacher Ratio  Year  Poverty Level  High Poverty  Strict Poverty  \
0                 16.2  1998              1             1               0
1                 17.6  1998              1             1               0
2                 18.5  1998              1             1               0
3                 16.9  1998              2             1               1
4                 16.6  1998              2             1               1

   No Poverty
0           0
1           0
2           0
3           0
4           0

[5 rows x 22 columns]
```

Poverty Level Threshold

- Poverty Level (CCD) is determined by percentage of students that qualify for Free/Reduced Lunch.
  - High (>75%)
  - Mid-high (50.1% - 75%)
  - Mid-low (25.1% - 50%)
  - Low (<25%)
- For our model we will take a similar approach:
  - Strict (>=66%)
  - High (>=33%)
  - No poverty (<33%)



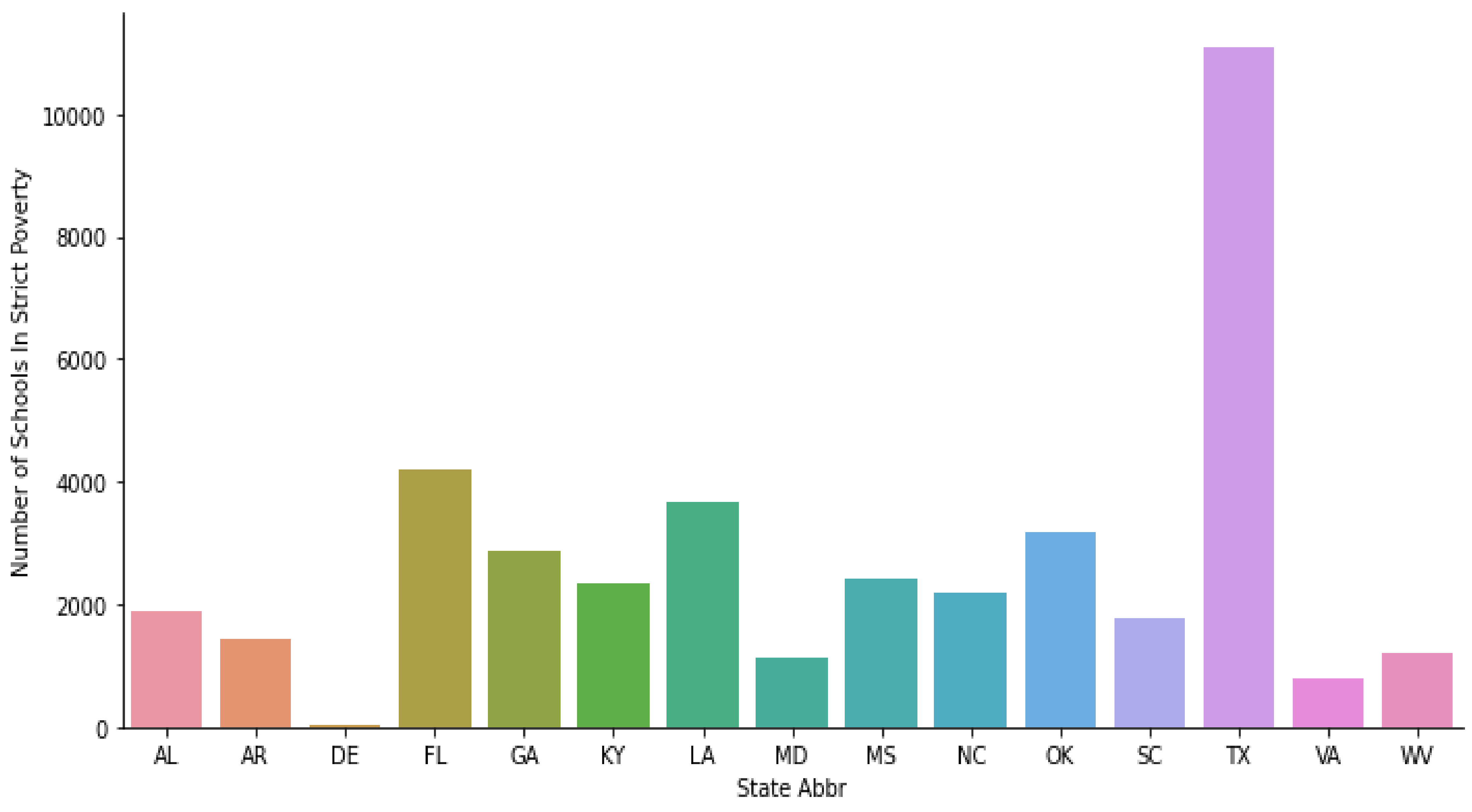
School Distribution Per State



State	# Schools
AL	4282
AR	4827
DE	531
FL	7045
GA	5813
KY	4542
LA	5302

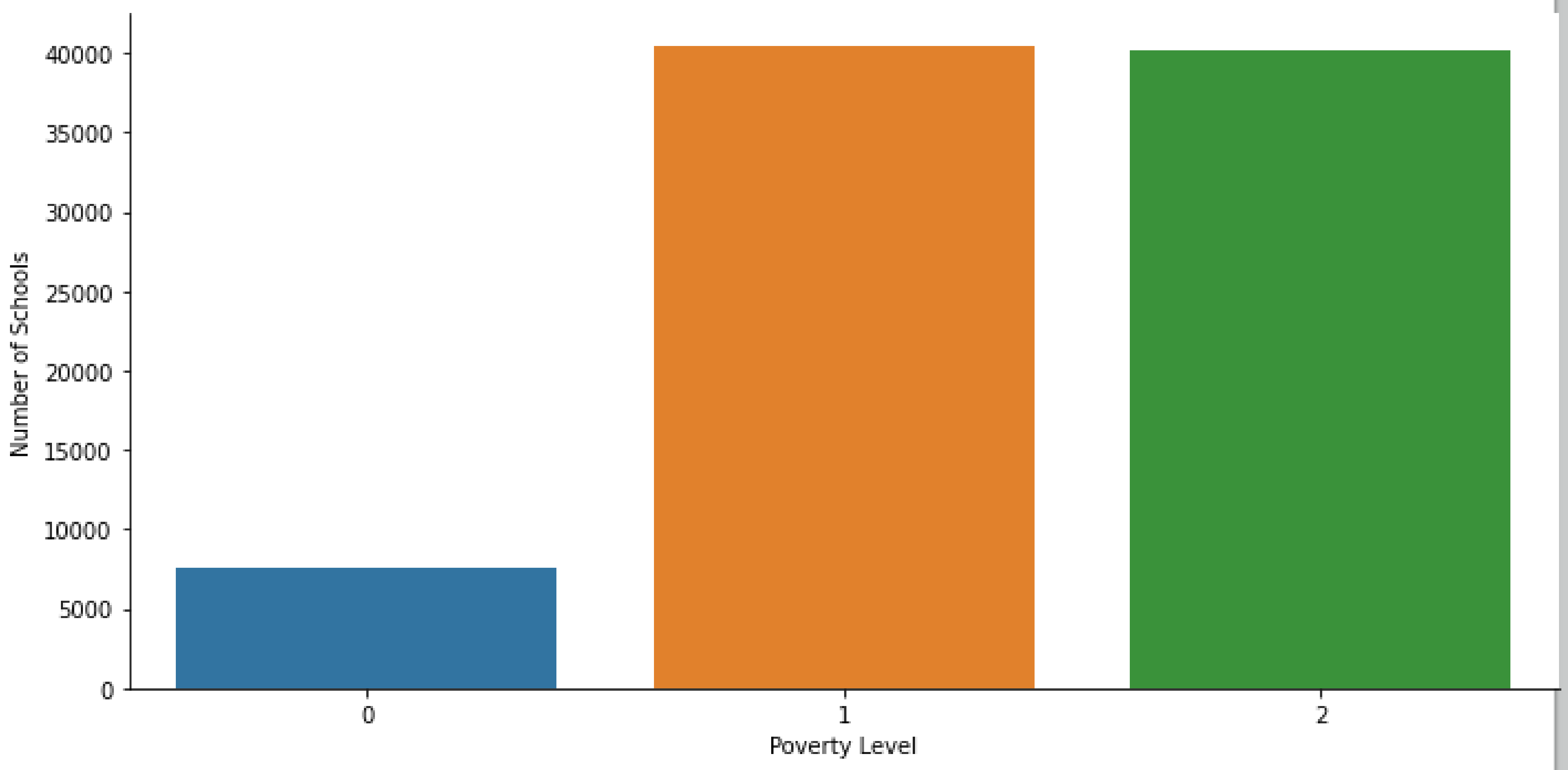
State	# Schools
MD	2559
MS	4023
NC	6212
OK	7026
SC	3098
TX	27177
VA	3156
WV	2565

Strict Poverty Per State



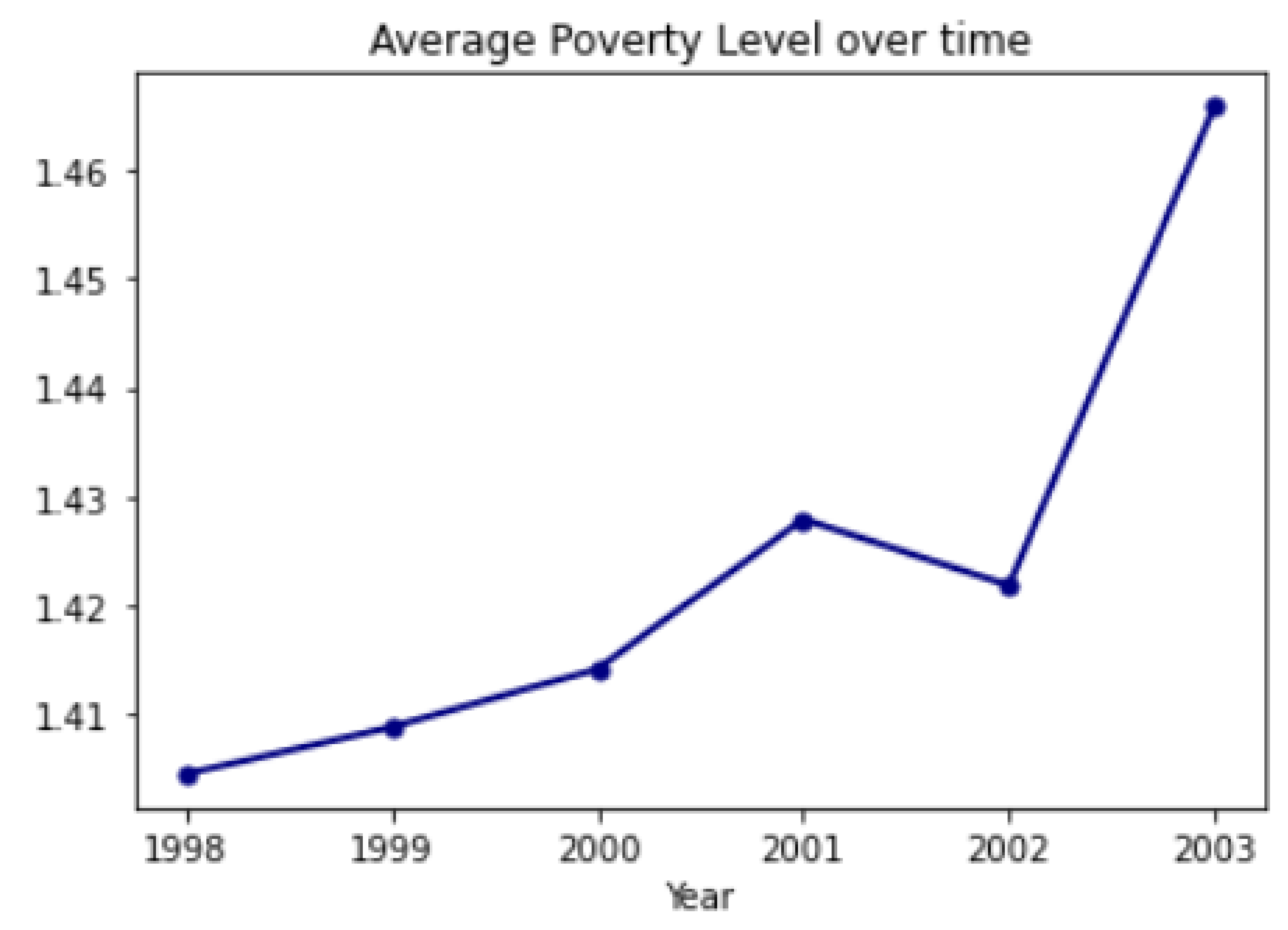
State	# Schools	State	# Schools
AL	1904	MD	1112
AR	1436	MS	2411
DE	31	NC	2186
FL	4214	OK	3185
GA	2858	SC	1767
KY	2331	TX	11088
LA	3676	VA	791
		WV	1198

Poverty Level vs Schools



Poverty Level	Number Of Schools
No Poverty	7551
High Poverty	40419
Strict Poverty	40188

Average Poverty vs Time



"Poverty Rates Fell in 2000, But Income Was Stagnant"

Feature Selection

- Years: 2004-2006
- Before Feature Selection begins, we can already eliminate the following columns:
  1. ~~School~~
  2. ~~State~~
  3. ~~State Abbr~~
  4. ~~School ID~~
  5. ~~Agency ID~~
- From our Poverty Level Threshold, we have added 4 new columns:
  1. No Poverty Level (0, 1)
  2. High Poverty Level (0, 1)
  3. Strict Poverty Level (0, 1)
  4. Poverty Level (0, 1, 2)

Feature Selection

- First, we will determine which features are most relevant for determining Poverty Level:
  1. For this we used a 'chi2' test on all features. The higher the score, the best the feature is at predicting poverty level.
     - Nominal 2-class target variable
     - Multiple dependent variables
  2. We also looked at our Correlation Matrix.
- Then, we establish the optimal number of features.

  To do this we used Forward Selection using a tree-based model ('ExtraTreesClassifier') to determine the ideal number of features to predict Poverty Level.
  - **3 is the optimal number of features**

Correlation Matrix

Based on the Correlation Matrix, the most significant features for both Title I and Poverty Levels are:

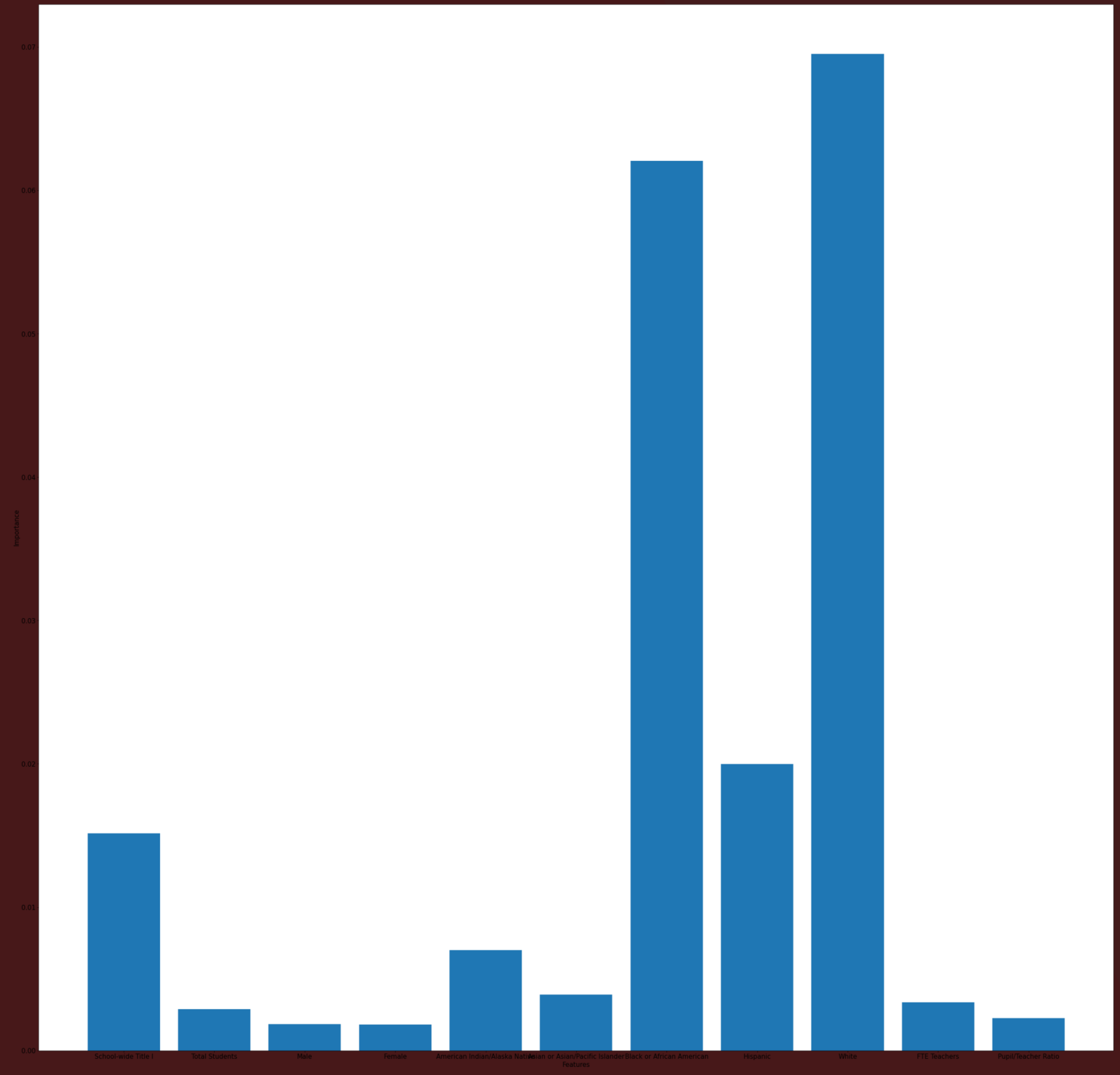
- Black or African American
- White
- Hispanic

Chi Square Scores(with Y = 'strict')

Specs	Chi square score
School-wide Title I	520.878645
Male	0.194143
Female	0.036934
American Indian/Alaska Native	57.565779
Asian or Asian/Pacific Islander	22.946501
Black or African American	3236.796141
Hispanic	988.992489
White	3146.773986
Pupil/Teacher Ratio	86.134895

Feature Selection Results

- Based on our results for ExtraTreesClassifier, Correlation Matrix, and Chi Square Scores, we determine that the 3 features we will use for modeling are:
  - Black or African American
  - White
  - Hispanic



Modeling

- Years: 2007-2015

LOGISTIC REGRESSION

- Compatible with classification problems
- Simple
- We need to make binary prediction

KNN CLASSIFIER

- Compatible with classifi cation problems
- No assumptions about data

SVM CLASIFIER

- Compatible with classi fication problems
- Linear SVC can be used for large sets of data

RANDOM FOREST

- Compatible with classi fication problems
- Works efficiently on large datasets
- Better accuracy than other classification models but more complex

Modeling Results

We made models for each of 'Title I', 'High Poverty Level', and 'Strict Poverty Level'. We evaluated each model by checking for accuracy and used 5-fold Cross Validation. These are the accuracy results on the Modeling dataset using the average 5-fold CV score.

LOGISTIC REGRESSION

- Title I : 92.13%
- High Poverty: 95.19%
- Strict Poverty: 72.68%

KNN CLASSIFIER (K=[1,2,…,10])

- Best Results: K=10
- Title I : 92.26%
- High Poverty: 95.37%
- Strict Poverty: 77.63%

SVM CLASIFIER (kernel=linear)

- Title I : 92.13%
- High Poverty: 95.19%
- Strict Poverty: 72.66%

RANDOM FOREST (n=100)

- Title I : 99.87%
- High Poverty: 99.67%
- Strict Poverty: 99.24%

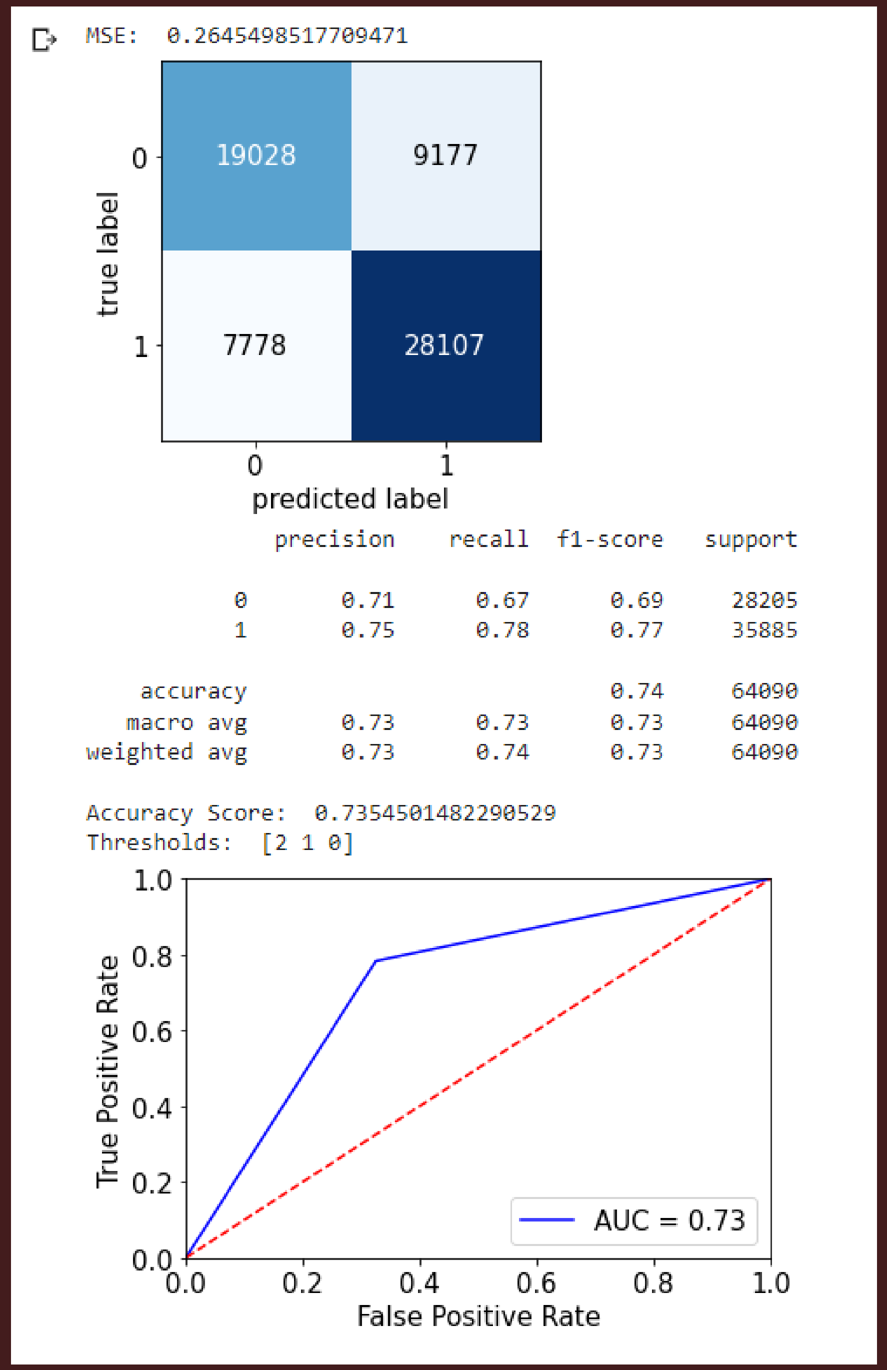
Model Selection

- Years 2015-2018
- For Model Selection we used the models that we built on the Modeling section and ran them with the Model Selection dataset.
- To evaluate our results, made a function evaluate(). This function automated the evaluation process of our model.
- The metrics used to evaluate are:
  - Mean Squared Error
  - Accuracy, Precision, Recall, F1 Score
  - Confusion Matrix
  - ROC-AUC Curve

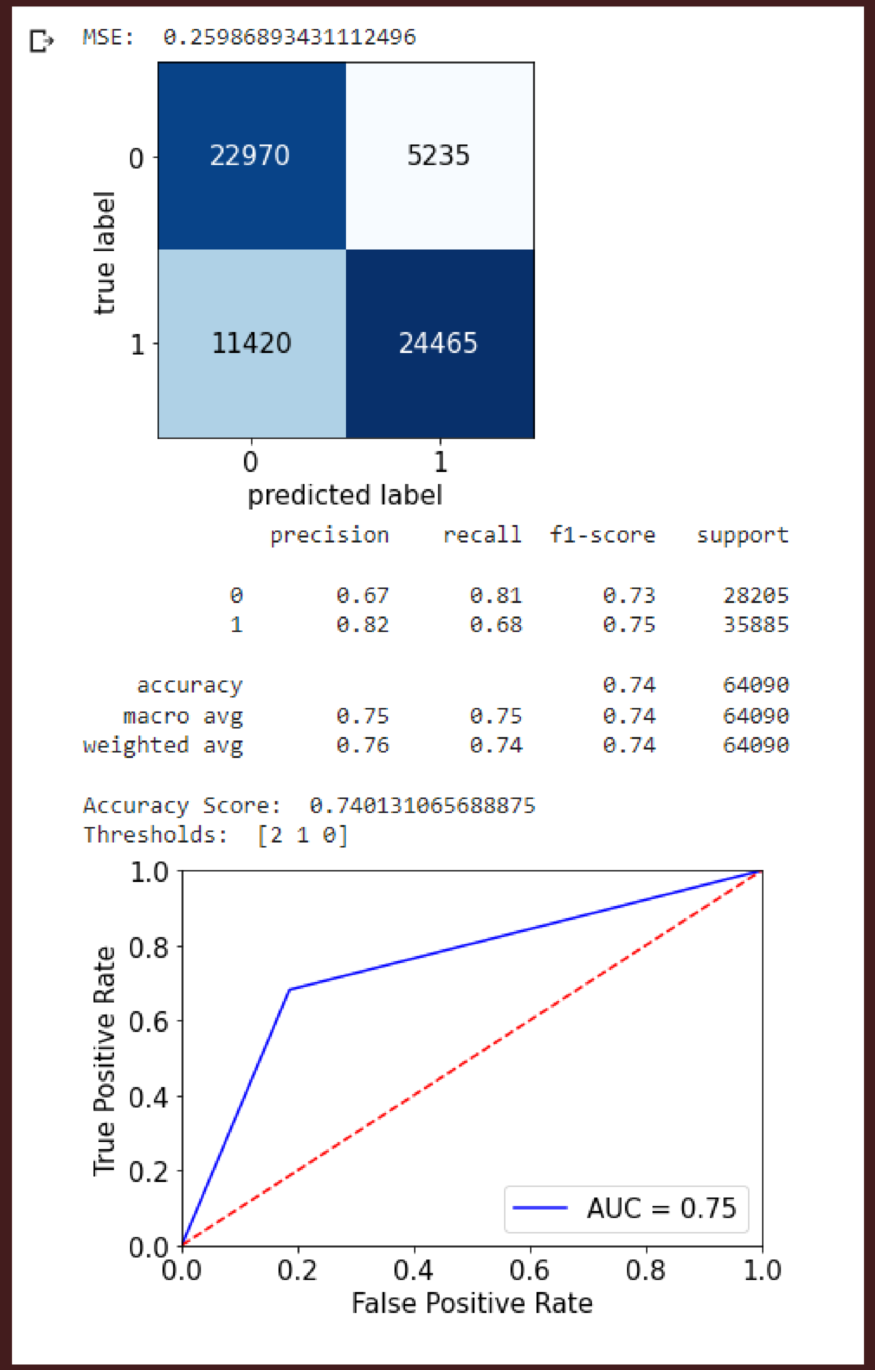
```
def evaluate(clf, X, y, cv=5):
    y_pred = clf.predict(X)

    print('MSE: ', mean_squared_error(y, y_pred))

    cm = confusion_matrix(y, y_pred)
    fig, ax = plot_confusion_matrix(cm)
    plt.show()

    print(classification_report(y, y_pred, target_names=['Strict Poverty', 'Not Strict Poverty']))

    fpr, tpr, thresholds = roc_curve(y, y_pred)
    roc_auc = auc(fpr, tpr)
    print('Thresholds: ', thresholds)
    plt.plot(fpr, tpr, 'b', label = 'AUC = %0.2f' % roc_auc)
    plt.legend(loc = 'lower right')
    plt.plot([0, 1], [0, 1],'r--')
    plt.xlim([0, 1])
    plt.ylim([0, 1])
    plt.ylabel('True Positive Rate')
    plt.xlabel('False Positive Rate')
    plt.show()
```

Model Selection Results (Strict Poverty)

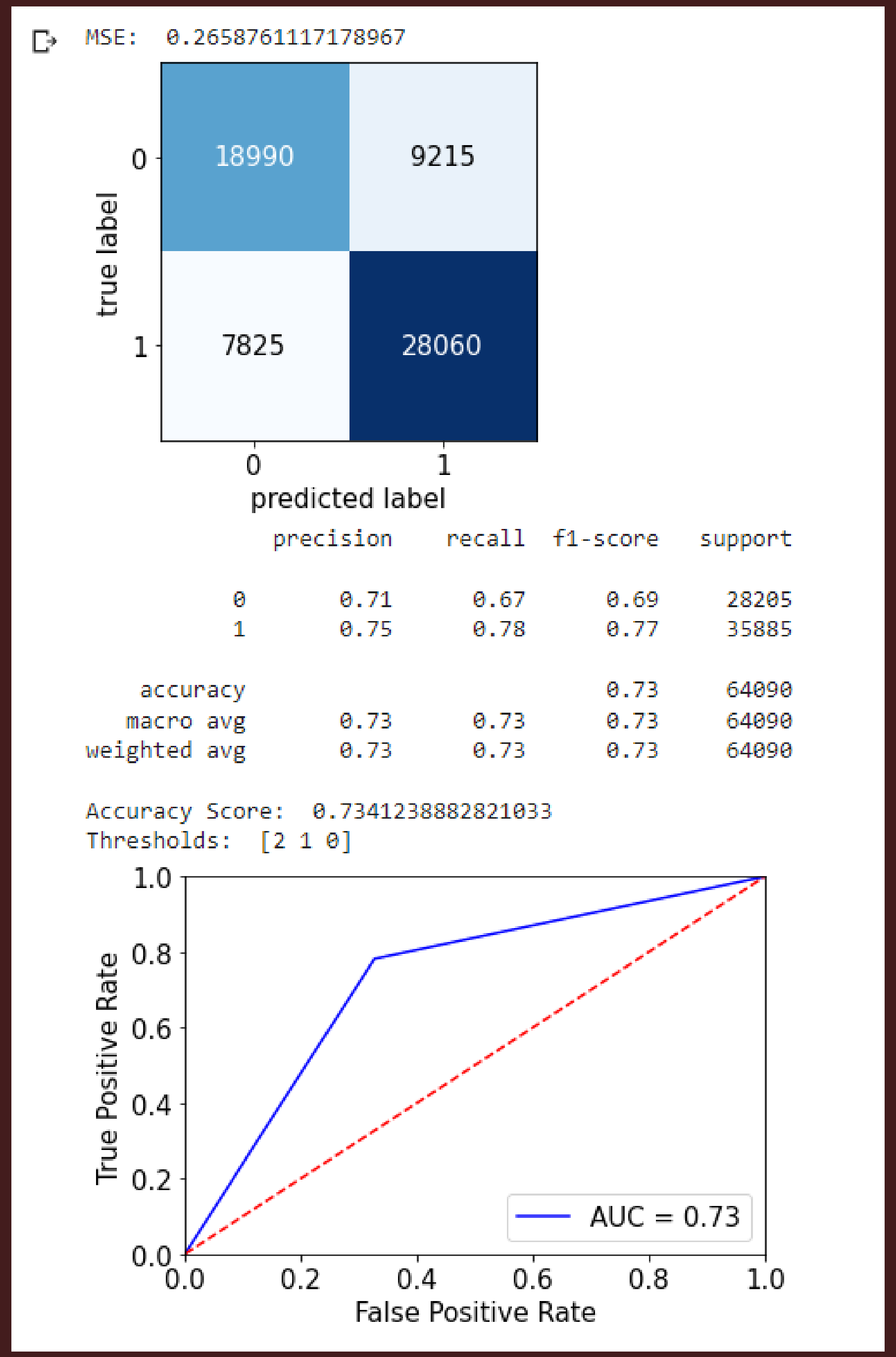
LOGISTIC REGRESSION



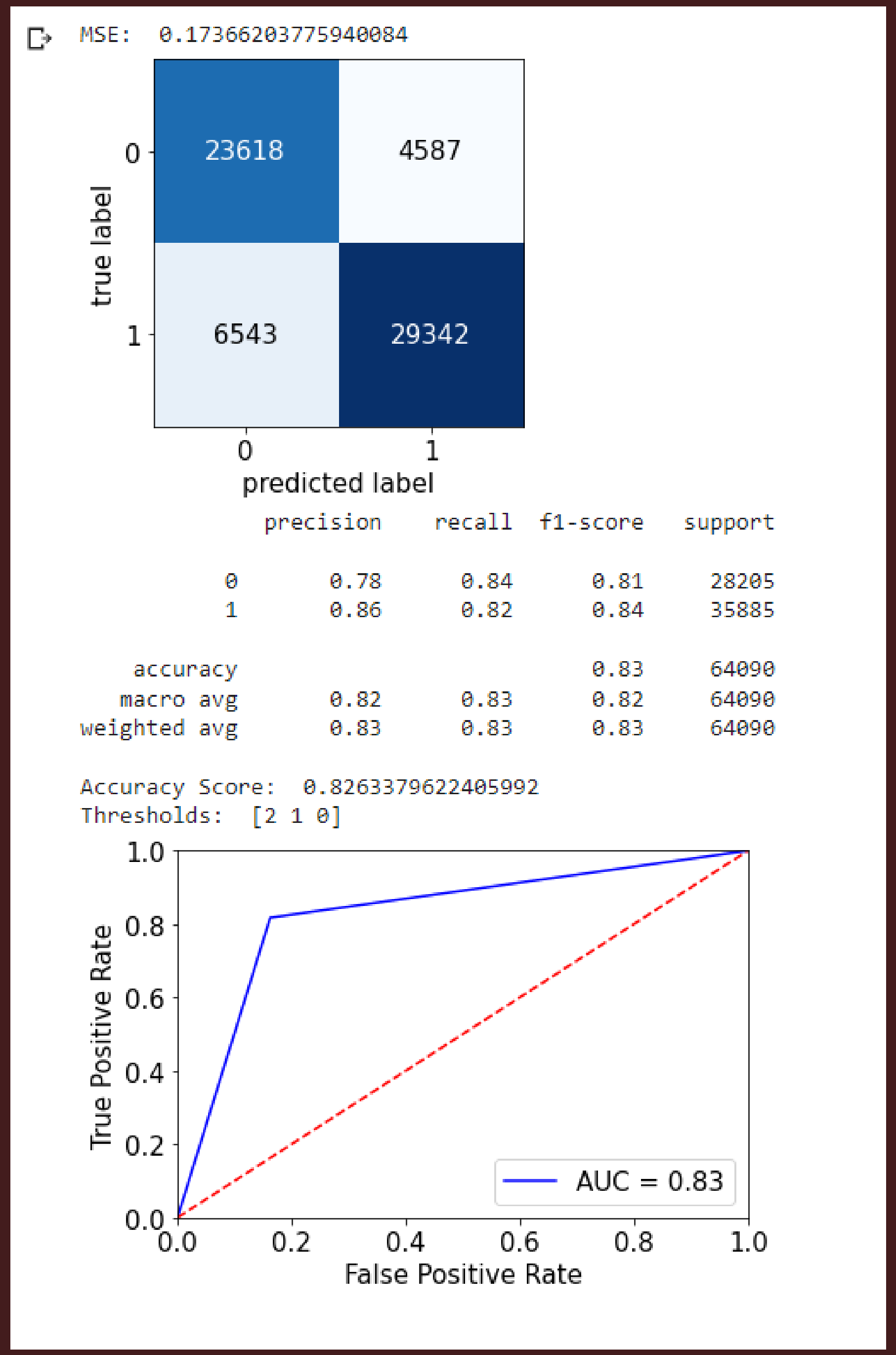
KNN CLASSIFIER



SVM CLASIFIER

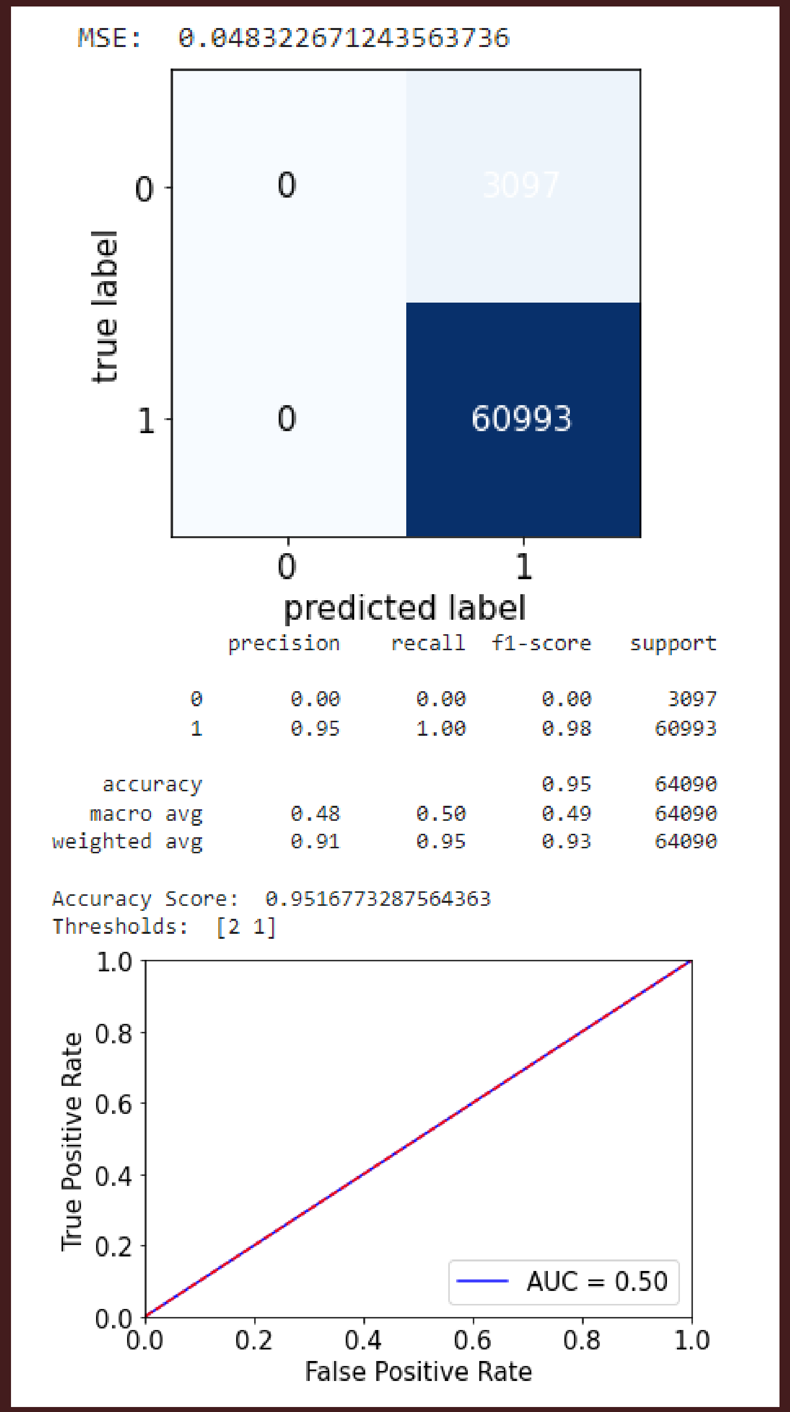


RANDOM FOREST

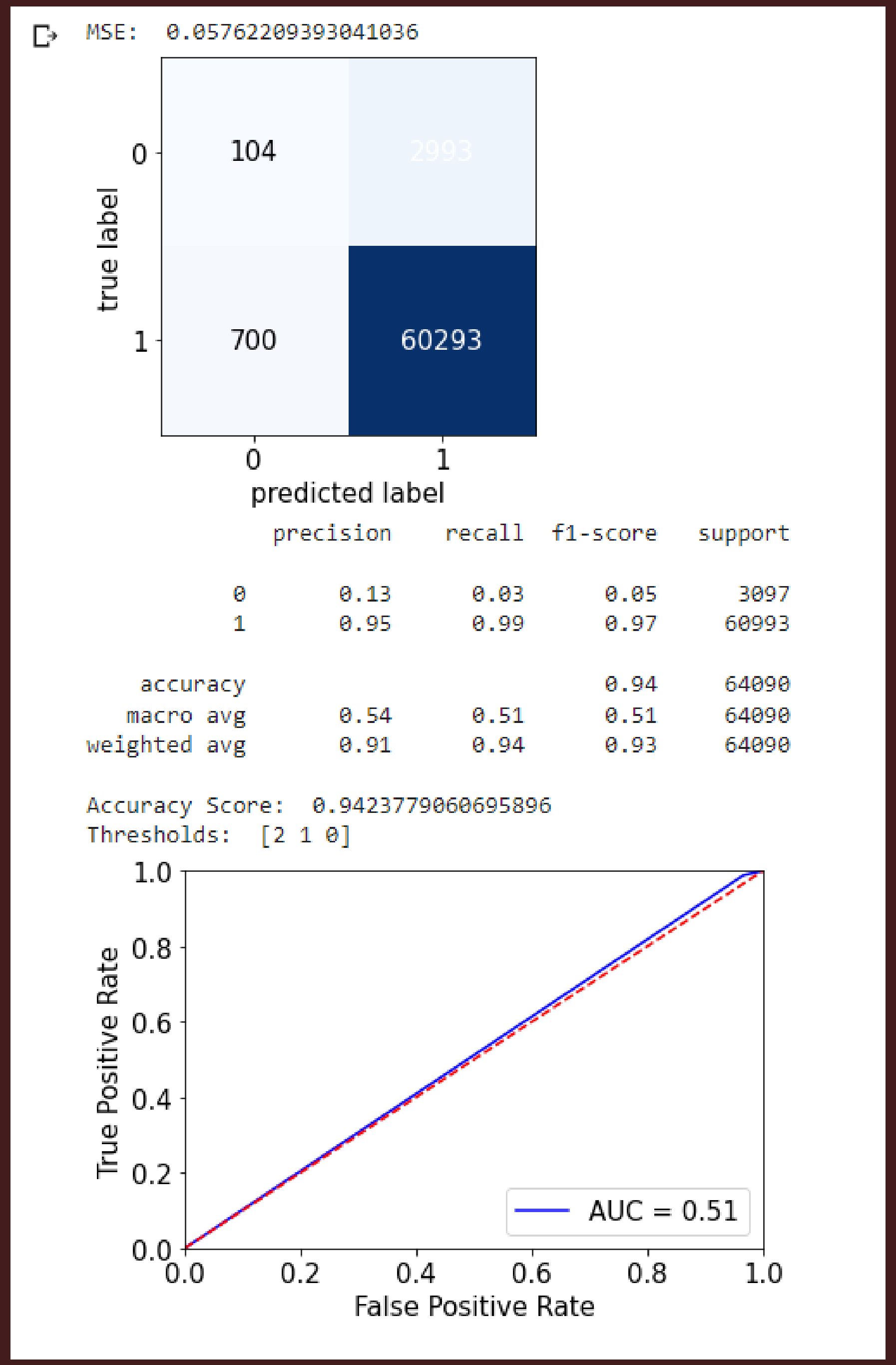


Model Selection Results (High Poverty)

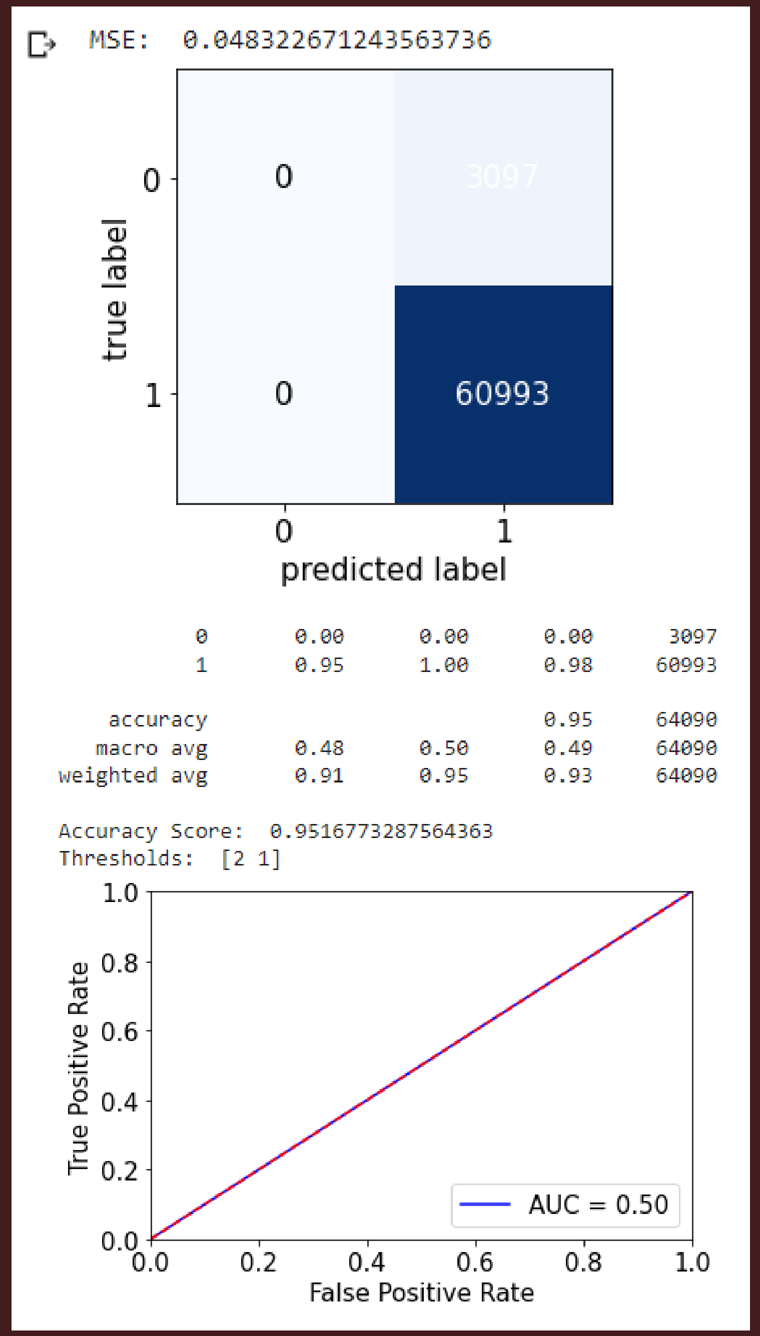
LOGISTIC REGRESSION



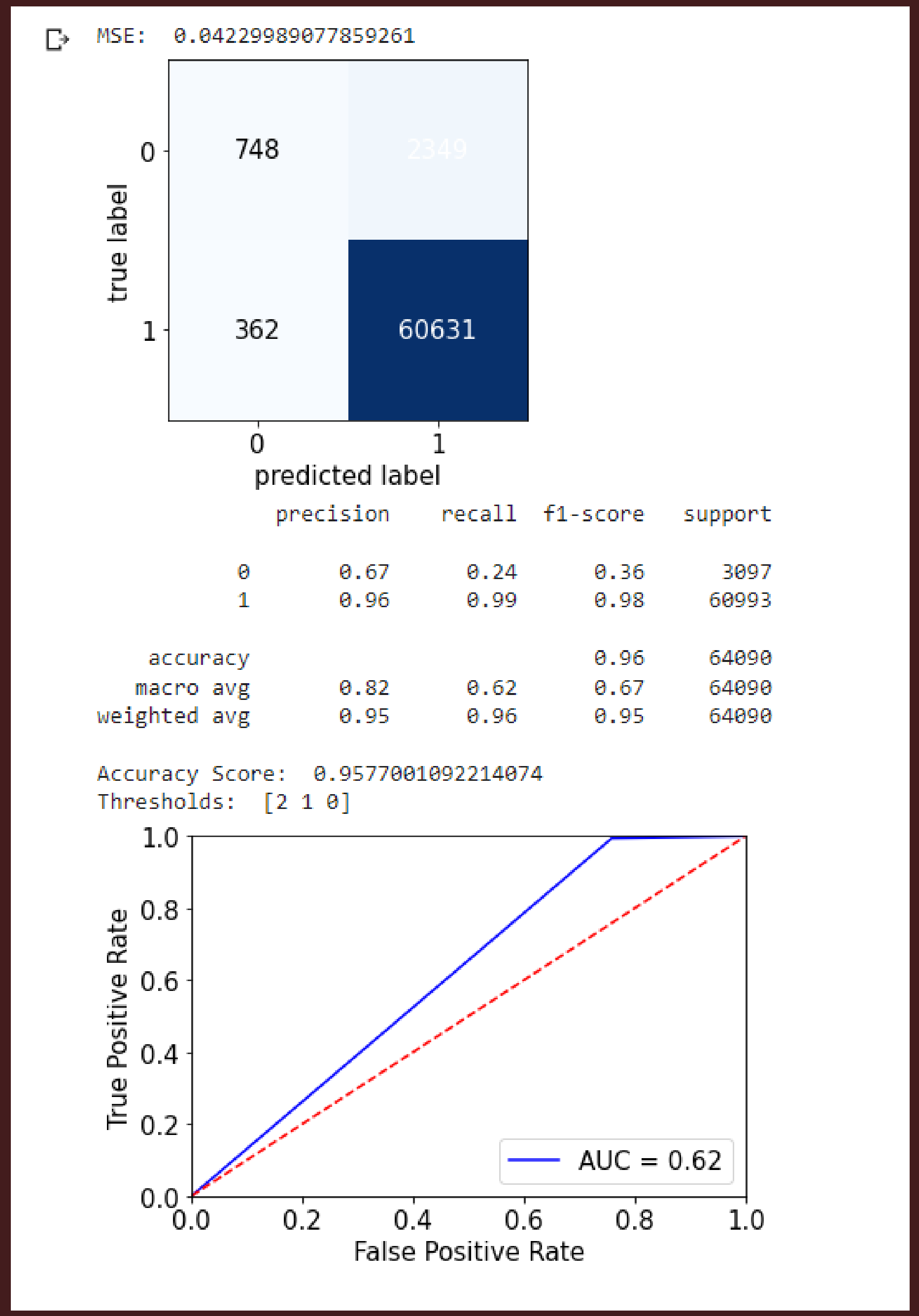
KNN CLASSIFIER



SVM CLASIFIER

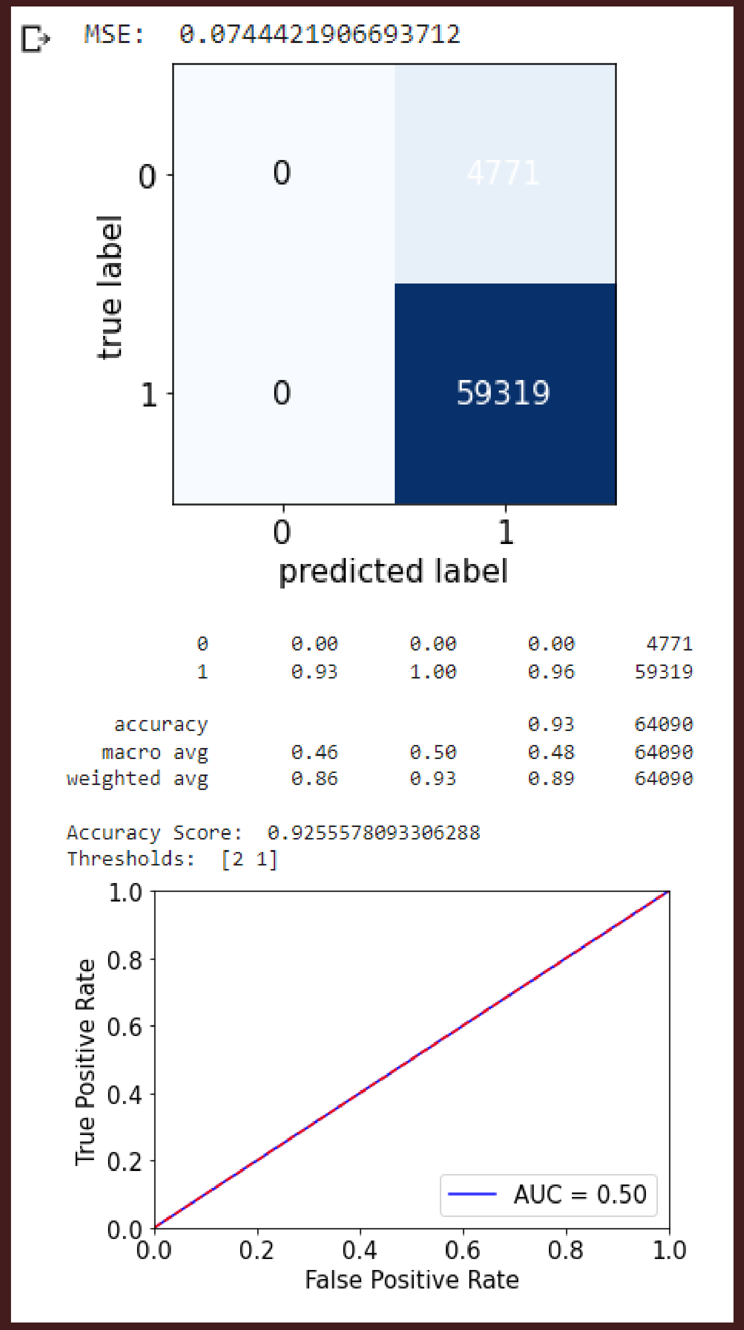


RANDOM FOREST

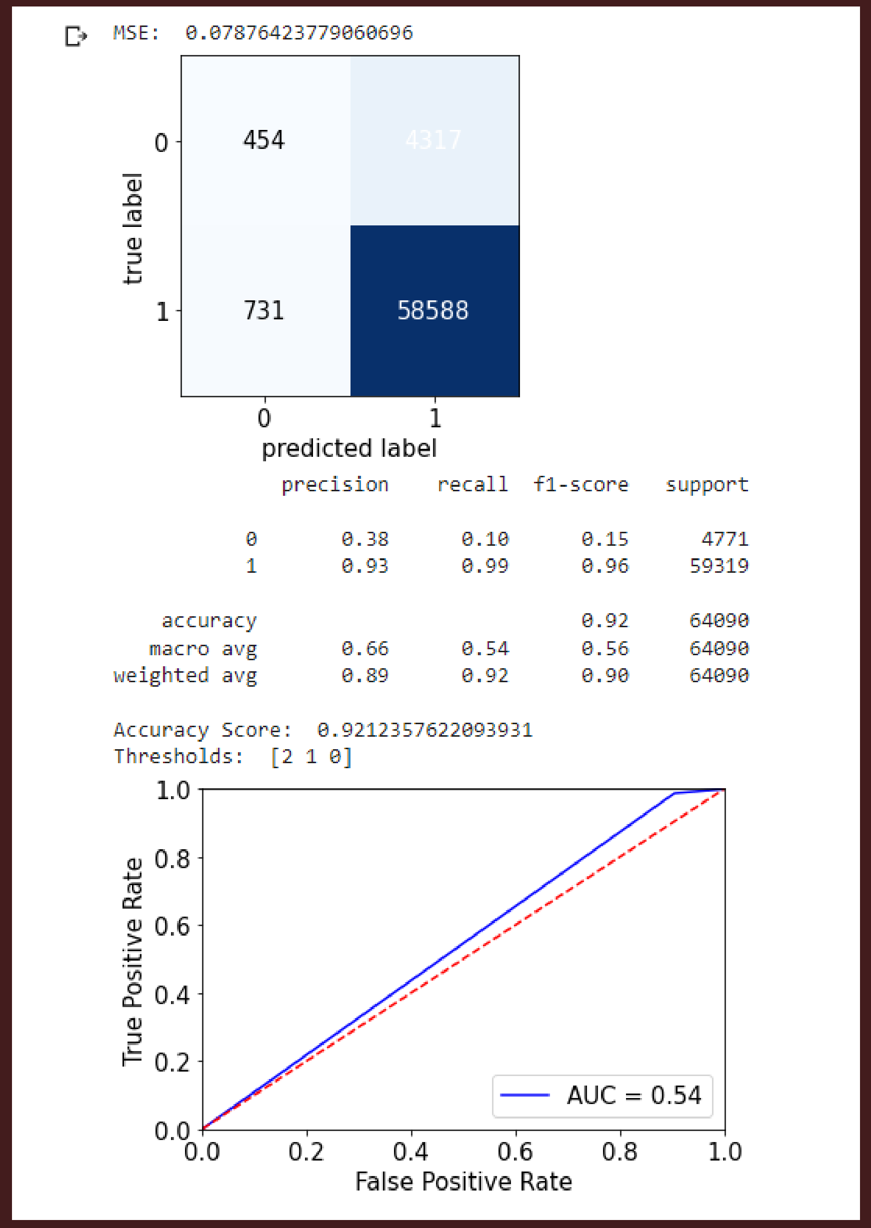


Model Selection Results (School-wide Title I)

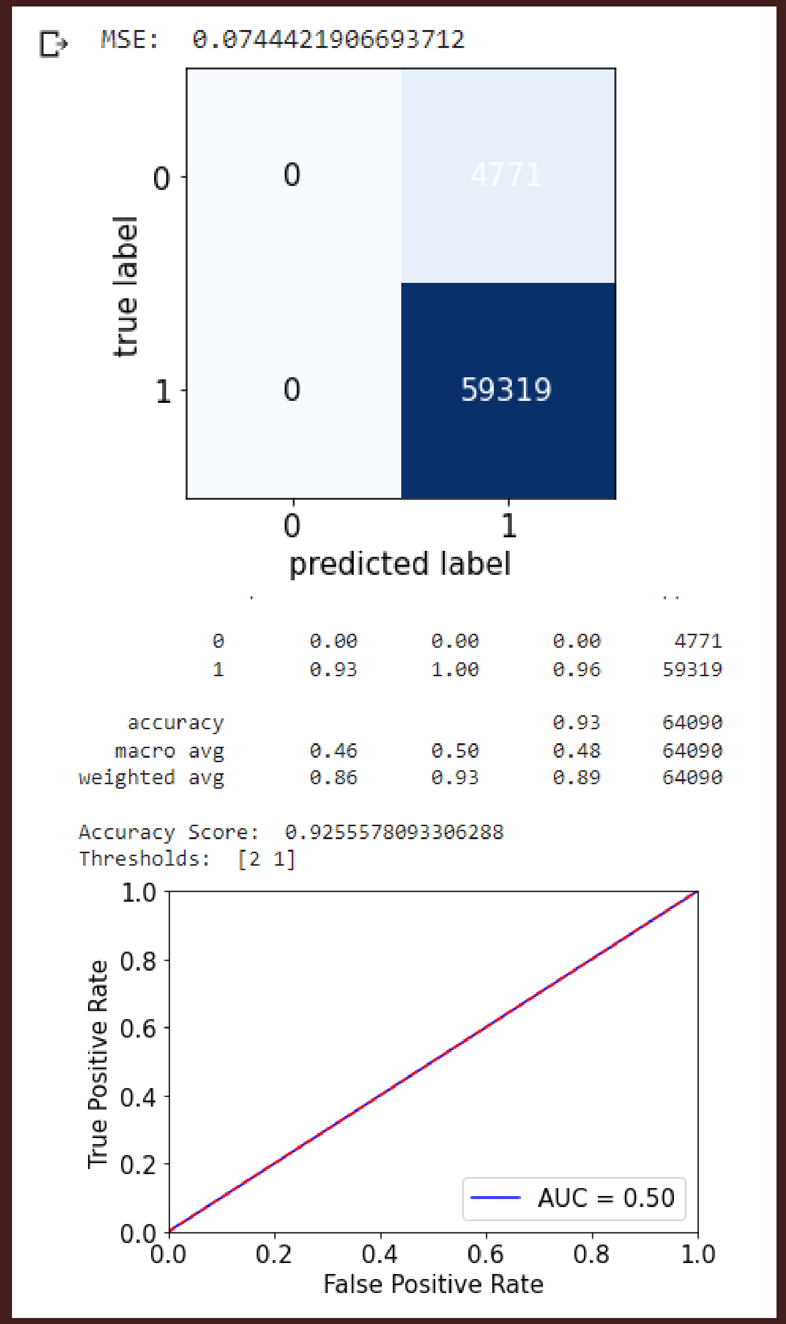
LOGISTIC REGRESSION



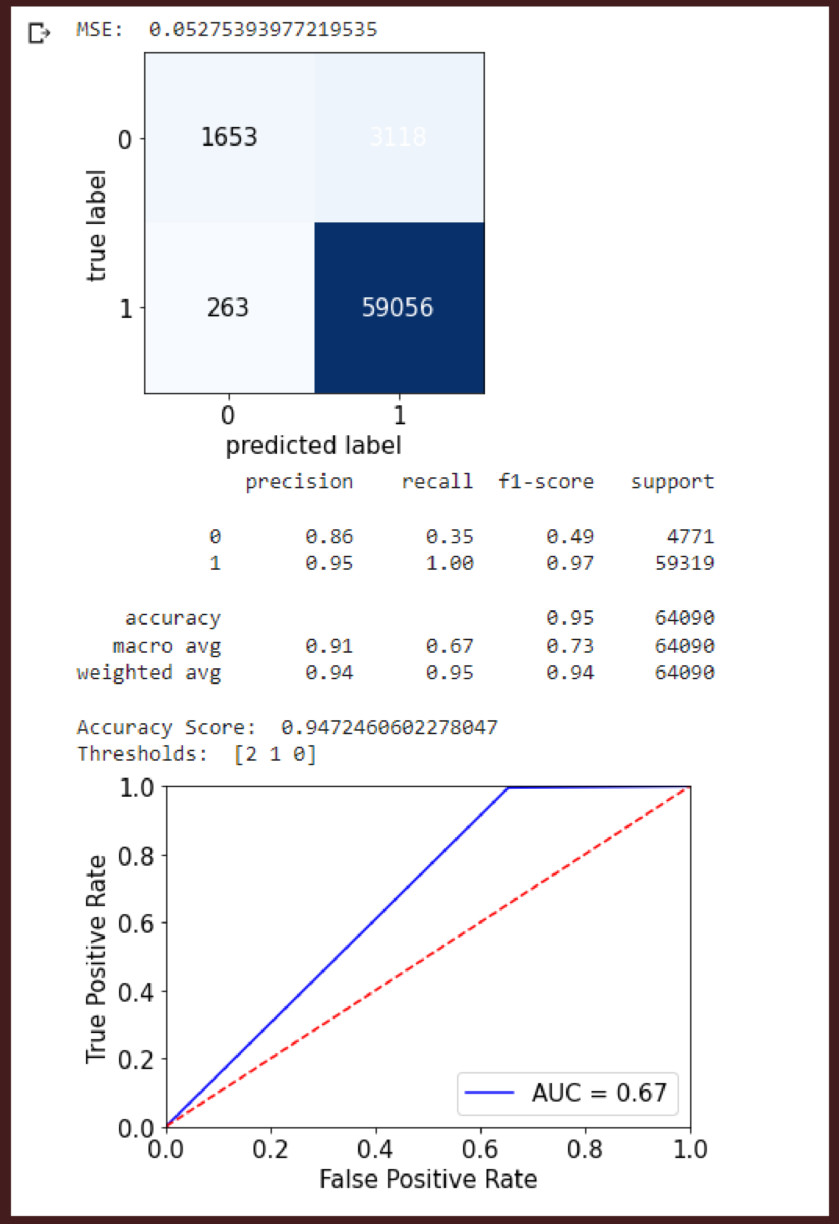
KNN CLASSIFIER



SVM CLASIFIER



RANDOM FOREST



Testing Results

- Years 2018-2020
- We have selected Random Forest to be the best classifier for our data.
- The Mean Squared Error for Random Forest is:

Strict Poverty Prediction Errors	High Poverty Prediction Errors	School-wide Title I Prediction Errors
27.21%	4.40%	5.56%

- Look for ways to reduce overfitting and improve regularization
- Revisit threshold for poverty level.