Logistic regression with many variables including state and county

A black line graph with white text

Description automatically generated

AUC 0.961

Confusion matrix from Logit Model

FALSE TRUE

0 58806 2149

1 284 6304

we got very good AUC and roc curve, also good predictions in the first model with state and county and every other variable i mentioned

but the AIC is much higher than your basic model, which meant that my first model is very bad

even though the logistic regression said all variables are significant

the odds ratio gave 0s and inf indicating the small variables are useless and some variabels gives perfect separation.

this perfect separation may because the lila is derived from lowincome. multicolinearity problem.

if lila is derived from lowincome we can't use it, its like giving the answers for an exam

#### Stepwise model results

This gave us good AIC and less deviance. The coeffiicients are not fully 0 and inf. Lowincome alone is inf as expected. We should tell this to darcy. How we found the derived variables from analysis.

variables like "PCTGQTRS", medianfamilyincome are not present.

we have something to show prof, how the straightforward logistic is bad and we need to do stepwise model specifically for prediction

variables used   
c("Urban","GroupQuartersFlag","LowIncomeTracts","lahunv1share","PCTGQTRS","MedianFamilyIncome","lawhite1","lablack1","laasian1","lahisp1","lanhopi1","laomultir1","laaian1","lakids10","lakids1","TractKids","laseniors1","laseniors10","TractKids","TractSeniors","TractWhite","TractBlack","TractAsian","TractNHOPI","TractAIAN","TractOMultir","TractHispanic","TractHUNV","TractSNAP", "PovertyRate")

variables got

Call:

glm(formula = LILATracts\_1And10 ~ LowIncomeTracts + lahunv1share +

Urban + lakids10 + lawhite1 + laseniors10 + lablack1 + laomultir1 +

laasian1 + TractAsian + laaian1 + TractHUNV + TractWhite +

lakids1 + TractOMultir + GroupQuartersFlag + PovertyRate +

TractSNAP, family = binomial(link = "logit"), data = cleaned\_data)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.19e+07 3.33e+05 -35.83 < 2e-16 \*

LowIncomeTracts1 1.19e+07 3.33e+05 35.83 < 2e-16 \*

lahunv1share 5.47e+01 3.84e+00 14.23 < 2e-16 \*

Urban1 7.48e+01 2.23e+00 33.58 < 2e-16 \*

lakids10 2.74e-01 1.25e-02 21.89 < 2e-16 \*

lawhite1 6.06e-03 2.91e-04 20.86 < 2e-16 \*

laseniors10 3.30e-01 1.17e-02 28.14 < 2e-16 \*

lablack1 7.83e-03 4.03e-04 19.43 < 2e-16 \*

laomultir1 1.05e-02 7.31e-04 14.34 < 2e-16 \*

laasian1 3.20e-02 3.16e-03 10.14 < 2e-16 \*

TractAsian -2.19e-03 3.66e-04 -5.98 2.2e-09 \*

laaian1 7.28e-03 1.02e-03 7.14 9.3e-13 \*

TractHUNV -3.86e-03 4.52e-04 -8.54 < 2e-16 \*

TractWhite 2.31e-04 3.77e-05 6.13 8.8e-10 \*

lakids1 -3.71e-03 1.04e-03 -3.55 0.00038 \*

TractOMultir -3.28e-04 8.75e-05 -3.75 0.00017 \*

GroupQuartersFlag1 1.85e+00 3.68e-01 5.02 5.1e-07 \*

PovertyRate -1.85e-02 3.93e-03 -4.71 2.5e-06 \*

TractSNAP 1.35e-03 3.08e-04 4.38 1.2e-05 \*

---

Signif. codes: 0 ‘\*’ 0.001 ‘\*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 43178.8 on 67542 degrees of freedom

Residual deviance: 4248.7 on 67524 degrees of freedom

AIC: 4287

Number of Fisher Scoring iterations: 25

###### Odd ratio for stepwise function

A screenshot of a computer

Description automatically generated

(Intercept) 0.0000e+00: The odds ratio for the intercept is not typically interpreted in isolation. It represents the odds of the outcome when all predictor variables are zero (which may not always be meaningful, especially if zero is not within the range of all predictors).

LowIncomeTracts1 Inf: An odds ratio of infinity (Inf) suggests that the presence of this variable (assuming it's a binary indicator) vastly increases the odds of the outcome, potentially to an unlimited extent. This could indicate a case of perfect separation, where the outcome is always 1 when this predictor is present.

lahunv1share 5.6307e+23: This extremely large odds ratio suggests that a one-unit increase in lahunv1share is associated with a substantial increase in the odds of the outcome. However, such a large value might indicate issues with the model, such as overfitting or separation.

Urban1 2.9661e+32: Similarly, this indicates a very large increase in the odds of the outcome for urban areas (assuming Urban1 is a binary indicator for urban areas). Again, such a large value might not be practically interpretable.

lakids10 1.3158e+00: This suggests that a one-unit increase in lakids10 is associated with about a 31.58% increase in the odds of the outcome.

lawhite1 1.0061e+00: A one-unit increase in lawhite1 is associated with approximately a 0.61% increase in the odds of the outcome.

laseniors10 1.3911e+00: This indicates that a one-unit increase in laseniors10 is associated with about a 39.11% increase in the odds of the outcome.

lablack1, laomultir1, laasian1, TractAsian, laaian1, TractHUNV, TractWhite, lakids1, TractOMultir, PovertyRate, TractSNAP: Each of these variables has an odds ratio close to 1, suggesting only a minimal change in the odds of the outcome per unit increase in these variables.

GroupQuartersFlag1 6.3374e+00: This indicates that being in a group quarters (assuming this variable indicates that) is associated with about a 533.74% increase in the odds of the outcome.

Important Considerations:

Context: The interpretation of each variable should be in the context of the specific data and the domain from which it comes. For example, understanding what lakids10 represents is crucial to interpret its effect meaningfully.

Magnitude: Extremely large values, especially Inf, should be interpreted with caution. They often indicate issues with the data or model, such as perfect separation.

Binary vs Continuous Variables: The interpretation differs for binary predictors (like Urban1, which might be a 0/1 indicator) and continuous predictors. For binary predictors, the odds ratio compares the odds of the outcome for one category against the other. For continuous predictors, it's about the change in odds per unit increase in the predictor.

Statistical Significance: An odds ratio should be interpreted in conjunction with its confidence interval and p-value. A large odds ratio with a wide confidence interval or a high p-value might not be significant.

Model Fit and Assumptions: Ensure the model is a good fit for your data and that the assumptions of logistic regression are met before heavily interpreting these results.