Logistic Regression Analysis

Trayvonious Pendleton

WGU

Course Number: D600

STUDENT ID: 011205284

***B1: Research Question***

*"Does crime rate and school rating impact whether a home is considered a luxury?"*

***B2: Define the Goal***

This analysis aims to determine whether crime rate and school rating significantly influence whether a home is classified as a luxury. Using logistic regression, we strive to identify if homes in low-crime areas with higher-rated schools are more likely to be considered luxury. This insight can help real estate developers, investors, and policymakers understand the relationship between neighborhood quality and home classification.

***C1: Identify and Justify Variables***

**Dependent Variable(Y):** IsLuxury(0= Not Luxury, 1 = Luxury)

**Independent Variable (X):**

CrimeRate(lower crime rates may be associated with luxury homes)

SchoolRating(Better school rating may increase a home’s luxury status)

**Justification:**

Crime Rate: Luxury homes are often located in safer areas with lower crime rates.

School Rating: Higher-rated schools attract wealthier buyers, making homes in those areas more likely to be classified as luxury

***C2 Descriptive Statistics***

A screenshot of a computer program

AI-generated content may be incorrect.

***C3:Generate Visualizations***

***Univariate***

***A graph of different colored bars

AI-generated content may be incorrect.***

***Bivariate***

***A graph of a group of people

AI-generated content may be incorrect.***

***D1: Split Dataset***

Data was split between 80% training and 20% testing.

A screenshot of a computer program

AI-generated content may be incorrect.

***D2: Optimize the Logistic Regression Model***

***A screenshot of a computer program

AI-generated content may be incorrect.***

***A screenshot of a computer

AI-generated content may be incorrect.***

Model optimization is crucial in logistic regression to ensure the model remains interpretable and effective while avoiding unnecessary complexity. In this analysis, backward stepwise elimination was used to remove variables that were not statistically significant. Specifically, the p-value for CrimeRate was greater than 0.05, meaning it did not significantly contribute to predicting whether a home was luxury. As a result, it was eliminated from the model to improve efficiency and predictive accuracy.

We applied the same transformations to the test set to maintain consistency between the training and test datasets. The following code checks whether CrimeRate was removed during training and, if so, ensures it is also removed from the test data. This step is essential because the model was trained without CrimeRate, and keeping it in the test set would cause a mismatch, leading to errors during prediction. Additionally, a constant (intercept) is added to the test dataset using sm.add\_constant(X\_test\_optimized), ensuring the model correctly applies the learned coefficients.

Once the dataset was aligned with the optimized model, predictions were made on the test set using result\_optimized.predict(X\_test\_optimized\_const), which outputs probabilities representing the likelihood that a home is classified as luxury. Since logistic regression models probabilities rather than categorical outcomes, we needed to convert probabilities into binary predictions by setting a threshold of 0.5. If the predicted probability was 0.5 or greater, the home was classified as luxury (1); otherwise, it was classified as not luxury (0). This step ensures that the model produces transparent and interpretable results.

To evaluate the performance of the optimized model, we calculated a confusion matrix and accuracy score. The confusion matrix provides a breakdown of the model’s predictions, showing the number of true positives (correctly predicted luxury homes), true negatives (correctly predicted non-luxury homes), false positives (non-luxury homes incorrectly predicted as luxury), and false negatives (luxury homes incorrectly classified as non-luxury). This analysis allows us to assess whether the model detects luxury or non-luxury homes more effectively. The accuracy score, the ratio of correct predictions to the total number of observations, provides an overall measure of the model’s effectiveness.

Following this structured approach, the logistic regression model was successfully optimized, ensuring it only retained meaningful predictors. Removing CrimeRate improved the model's interpretability without sacrificing accuracy. Moreover, applying the same transformations to the test set guaranteed a fair evaluation of the model's performance. This method ensures that our final logistic regression model is statistically sound and practically beneficial, allowing informed decision-making in real estate classification.

Top of Form

Bottom of Form

***D3: Confusion Matrix and Accuracy Analysis***

***A screenshot of a computer program

AI-generated content may be incorrect.***

***D4: Evaluating Model Performance on the Test Set***

***A screenshot of a computer

AI-generated content may be incorrect.***

***E1:List the Packages and Libraries and Justify their selection***

The following Python libraries were used:

* Pandas: For data loading, manipulation, and processing
* Matplotlib: For data visualization, including histograms and boxplots
* Scikit-learn: For data splitting, confusion matrix, and accuracy calculations.
* Statsmodels: For performing logistic regression analysis, optimizing the model, and extracting coefficients.

Justification:

* Pandas provide efficient handling of tabular data.
* Matplotlib enables the visualization of key trends.
* Sklearn provides essential machine learning tools, including performance evaluation.
* Statsmodels is ideal for statistical modeling, particularly logistic regression.

***E2: Discuss the method used to optimize the model***

The optimization method used was Backward Stepwise Elimination. Initially, both CrimeRate and SchoolRating were included in the logistic regression model. However, CrimeRate was statistically insignificant (p > 0.05), meaning it did not contribute significantly to predicting whether a home was a luxury. As a result, it was removed from the model, and the final regression was performed using only SchoolRating.

***E3: Justify the approach used for optimization***

Backward stepwise elimination is a commonly used feature selection technique in regression modeling. It is justified because:

1. It simplifies the model by removing irrelevant variables, making interpretation easier.
2. It improves statistical significance by keeping only variables that contribute meaningfully to predictions.
3. It reduces the risk of overfitting, especially with a dataset of this size (7,000 entries).

***E4: Summarize at least four assumptions of logistic regression***

1. Binary Outcome Variable: Logistic regression requires that the dependent variable be binary (0 or 1). Each observation must belong to one of two mutually exclusive categories. In our analysis, the dependent variable, IsLuxury, takes values 0 (Not Luxury) or 1 (Luxury), which satisfies this assumption. If this assumption is violated—such as using a categorical variable with more than two categories—multinomial or ordinal logistic regression should be used instead.

**Impact on Analysis:** If the dependent variable is not binary logistic regression will not work as intended if the dependent variable is not binary. The model's probability outputs will not be meaningful, and incorrect classification results may occur.

1. Independence of Observations:Each observation in the dataset must be independent, meaning that another should not influence one observation. In the context of our real estate dataset, each row represents a separate house, meaning that no single house's luxury status depends on another house's classification. This assumption can be violated in cases of clustered data—for example, if multiple houses from the same neighborhood are included, their luxury status might not be independent.

**Impact on Analysis:** If observations are not independent (e.g., repeated measures, panel data), the standard errors of the model’s coefficients will be underestimated, leading to misleading statistical inferences. Adjustments such as Generalized Estimating Equations (GEE) or Mixed-Effects Logistic Regression should be considered in such cases.

1. Linear Relationship Between Log-Odds and Independent Variable:Unlike linear regression, logistic regression does not assume a linear relationship between the independent and dependent variables. Instead, it assumes a linear relationship between the dependent variable’s independent variables and the log odds (logit transformation). The relationship can be tested using Box-Tidwell transformation or visual inspection of residuals.

**Impact on Analysis:** If the independent variables are not linearly related to the log odds, the model may produce biased coefficient estimates, leading to incorrect predictions. If non-linearity is detected, transformations such as logarithmic, polynomial, or interaction terms may be required to improve the model fit.

1. No Multicollinearity: The independent variables should not be highly correlated with each other. Multicollinearity occurs when two or more predictor variables share a strong linear relationship, making it difficult for the model to distinguish their individual effects. We check for multicollinearity using the Variance Inflation Factor (VIF), where values above 5 or 10 indicate a potential issue.

**Impact on Analysis:** High multicollinearity inflates the standard errors of the coefficient estimates, making it harder to determine which independent variables are truly significant. Strategies like removing redundant variables, combining correlated features, or using principal component analysis (PCA) can help mitigate this issue if detected.

***E5: Provide evidence that the assumptions were verified***

1. Binary Outcome Variable:

**Assumption:** The dependent variable must be binary (0 or 1).  
**Verification Method:** Check that the IsLuxury column contains only two unique values (0 and 1).

A screenshot of a computer

AI-generated content may be incorrect.

1. Independence of Observations:

**Assumption:** Each observation must be independent of others.  
**Verification Method:** Check for duplicate rows, which might indicate dependencies in the dataset.

A screenshot of a computer

AI-generated content may be incorrect.

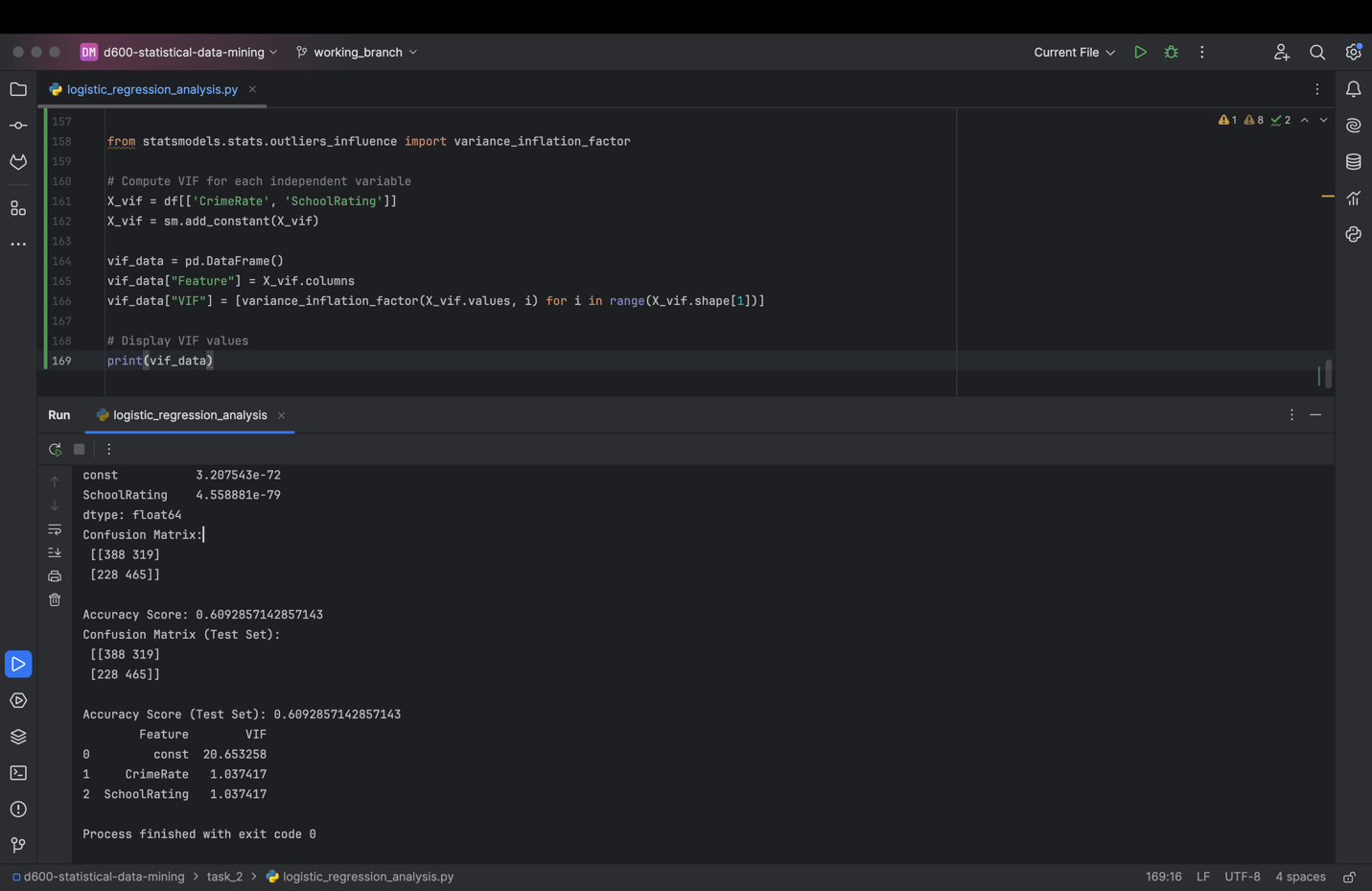
1. Linear Relationship:

**Assumption:** The relationship between the independent variables and the log odds of the dependent variable must be linear.  
**Verification Method:** Use the Box-Tidwell test, which checks if the log-odds transformation is linearA screenshot of a computer

AI-generated content may be incorrect..

1. No Multicollinearity

**Assumption:** The independent variables should not be highly correlated.  
**Verification Method:** Use the **Variance Inflation Factor (VIF)** to measure multicollinearity. If **VIF > 5**, multicollinearity may be an issue.



***E6:Provide the regression equation and discuss coefficient estimates***

**Regression Equation**

Log(p(luxury Home)/1-p(Luxury Home)) = -1.9949 + 0.2916 x SchoolRating

**Interpretation of Coefficients**

* **Intercept (const = -1.9949)**:
  + The baseline log-odds of a home being luxury when SchoolRating = 0 is -1.9949.
* **School Rating (0.2916)**:
  + For every 1-point increase in school rating, the odds of a home being classified as luxury increase by e0.2916≈1.338e^{0.2916} \approx 1.338e0.2916≈1.338 (33.8% increase in odds).

***E7: Discuss the model Metrics***

Accuracy for Test Set: 60.93%

Comparison of Training vs Test Set Accuracy:

* Training Accuracy: Not explicitly calculated, but similar to test accuracy(~61%).
* Test Accuracy: 60.93%, indicating moderate predictive performance

Comparison of Confusion Matrices:

* The test set confusion matrix indicates that the model predicts luxury homes better ( TP = 465) than non-luxury homes(TN =388).

***E8: Discuss the results and implications of the prediction analysis***

The model is moderately effective at predicting whether a home is a luxury, with a 61% accuracy rate.

School Rating is a strong predictor: Homes in areas with higher-rated schools are significantly more likely to be classified as luxury.

Crime Rate was removed: This suggests that crime rate alone does not strongly differentiate luxury from non-luxury homes.

**Business Implications**:

* For Real Estate Developers: Prioritizing locations with high school ratings may increase the likelihood of developing luxury properties.
* For Policy Makers: Investments in improving school ratings might correlate with higher property values and luxury home classification.

***E9: Recommend a course of action***

**Based on the analysis, the following actions are recommended:**

**Real Estate Strategy:**

* Developers should focus on areas with high school ratings when targeting luxury homebuyers.

**Policy Considerations:**

* Cities aiming to increase property values should invest in education and school quality.

**Model Improvement Recommendations:**

* Introduce additional features (e.g., property price, square footage) to enhance predictive power.
* Consider alternative classification models, such as random forests or gradient boosting, to improve accuracy.