Assignment

Logistic Regression

The banking industry in the USA is highly regulated. Banks are required to adhere to strict guidelines regarding the management of customer deposits, the maintaining of customer privacy, and the avoidance of discrimination when approving or rejecting loan applications. Many years ago, banks engaged in the practice of "redlining" when approving loans for home mortgages, which basically amounted to the automatic rejection of loan applications for homes located in certain areas that were deemed by banks to be overly prone to loan defaults. When such practices were deemed to be discriminatory by bank regulators, banks adjusted their mortgage lending practices in an attempt to make them more objective. However, regulators remain concerned that discrimination can exist even when what appear to be purely objective criteria are used for purposes of deciding whether or not a customer is approved for a loan. You've been tasked by a large banking regulator with the development of a model that can predict whether or not a given mortgage loan application is likely to be approved or denied. The regulator plans to use the output of such a model in an attempt to identify potential instances of discrimination in the lending practices of the banks for which it is tasked with regulating.

The data set you will be using is sourced from the Federal Reserve Bank of Boston. A brief overview of the content of the data set can be found here:

- https://vincentarelbundock.github.io/Rdatasets/doc/Ecdat/Hdma.html

The data set comprises 2,381 observations of 1 response/dependent variable (which indicates whether or not a mortgage application was denied) and 12 explanatory/independent variables. Please refer to the web page cited above for further details on these variables.

Your task for this Assignment is to construct and compare/contrast a series of **binary logistic regression models** (after completing the necessary EDA and data prep work) that predict whether or not a given mortgage application is likely to be denied. The response variable you will be modeling is the data set's "**DENY**" attribute, which indicates whether or not a mortgage application was denied. It is up to you as the data science practitioner to determine which features should be included in these models. Your work should include EDA, data preparation (including transforms as needed), feature selection, and a thorough evaluation of model performance metrics. Get started on the Assignment as follows:

1) Load the provided file to your Github Repository.
2) Then, using a Notebook, read the data set from your Github repository and load it into a Pandas dataframe.
3) Perform EDA work as necessary.
4) Perform any required data preparation work, including any feature engineering adjustments you deem necessary for your work.
5) Apply your knowledge of feature selection and/or dimensionality reduction techniques to identify explanatory variables for inclusion within your models. You may select the features manually via the application of domain knowledge, use forward or backward selection, or use a different feature selection method (e.g., decision trees, etc.). It is up to you as the data science practitioner to decide upon the most appropriate feature selection and/or dimensionality reduction techniques to be used

with the data set.

6) After splitting the data into training and testing subsets, use the training subset to construct at least three different binomial logistic regression models using different combinations of explanatory variables (or the same variables if they have been transformed via different transformation methods).

7) After training your various models, decide how you will select the "best" regression model from those you have constructed. For example, are you willing to select a model with slightly lower performance if it is easier to interpret or less complicated to implement? What metrics will you use to compare/contrast your models? Evaluate the performance of your models via cross validation using the training data set. Then apply your preferred model to the testing subset and assess how well it performs on that previously unseen data.

Your deliverable for this Assignment is your Jupyter Notebook. It should contain a combination of Python code cells and explanatory narratives contained within properly formatted Markdown cells. The Notebook should contain (at a minimum) the following sections (including the relevant Python code for each section):

1) **Introduction (5 Points)**: Summarize the problem + explain the steps you plan to take to address the problem

2) **Exploratory Data Analysis (20 Points)**: Explain + present your EDA work including any conclusions you draw from your analysis, including any preliminary predictive inferences. This section should include any Python code used for the EDA.

3) **Data Preparation (10 Points)**: Describe + show the steps you have taken to address the data integrity + usability issues you identified in your EDA, including any feature engineering techniques you have applied to the data set. This section should include any Python code used for Data Preparation.

4) **Prepped Data Review (5 Points)**: Explain + present your post-Data Prep EDA analysis. This section should include any Python code used for re-running your EDA on the variables adjusted during your Data Preparation work.

5) **Regression Modeling (40 Points)**: Explain + present your regression modeling work, including your feature selection work + interpretation of the coefficients your models are generating. Do they make sense intuitively? If so, why? If not, why not? Comment on the magnitude and direction of the coefficients + whether they are similar from model to model.

6) **Select Models (15 Points)**: Explain your model selection criteria. Identify your preferred model. Compare / contrast its performance with that of your other models. Discuss why you've selected that specific model as your preferred model. Apply your preferred model to the testing subset and discuss your results. Did your preferred model perform as well as expected?

7) Conclusions (5 Points)

Your Notebook deliverable should be similar to that of a publication-quality / professional caliber document and should include clearly labeled graphics, high-quality formatting, clearly defined section and subsection headers, and be free of spelling and grammar errors. Furthermore, your Python code should include succinct explanatory comments.