DAV 6150 Project 1 (Module 6)

Regression for Numeric Data

***** You may work in small groups of no more than four (4) people for this Project *****

For your first Project of the course you will be working with a dataset comprised of information pertaining to NY State High School graduation metrics for the 2018-2019 school year. The dataset is sourced from the NY State Education Department (NYSED): https://data.nysed.gov/downloads.php

The dataset is comprised of more than 73,000 observations, each of which pertains to a particular NY State school district and associated subgroupings/categorizations of high school students who had been enrolled for at least 4 years as of the end of the 2018-2019 school year. A data dictionary describing the attributes contained within the file is provided below.

Data Set Attribute	**Description**
report_school_year	Indicates school year for which high school graduation info is being reported
aggregation_index	Numeric code identifying manner in which high school graduation data has been aggregated
aggregation_type	Text description of how high school graduation data has been aggregated
nrc_code	Numeric code identifying "needs / resource capacity", which is an indicator of the type of school district
nrc_desc	Text description of the type of school district
county_code	Numeric code for county name
county_name	Full name of applicable NY State county
nyc_ind	Indicates whether or not the school district resides within the borders of NYC
membership_desc	Indicates school year in which students first enrolled in High School
subgroup_code	Numeric code identifying student subgrouping
subgroup_name	Text description of student subgrouping. Note that a student may belong to **MORE THAN ONE** subgrouping (e.g., "Female", "Hispanic", "Not English Language Learner", etc.)
enroll_cnt	How many students of the indicated subgrouping were enrolled during the given school year
grad_cnt	How many enrolled students of the indicated subgrouping graduated at the end of the given school year
grad_pct	What percentage of enrolled students of the indicated subgrouping graduated at the end for the given school year
local_cnt	How many enrolled students of the indicated subgrouping were awarded a "Local" diploma
local_pct	What percentage of enrolled students of the indicated subgrouping were awarded a "Local" diploma
reg_cnt	How many enrolled students of the indicated subgrouping were awarded a "Regents" diploma
reg_pct	What percentage of enrolled students of the indicated subgrouping were awarded a "Regents" diploma
reg_adv_cnt	How many enrolled students of the indicated subgrouping were awarded a "Regents Advanced" diploma
reg_adv_pct	What percentage of enrolled students of the indicated subgrouping were awarded a "Regents Advanced" diploma

non_diploma_credential_cnt	How many enrolled students of the indicated subgrouping achieved a non-diploma credential
non_diploma_credential_pct	What percentage of enrolled students of the indicated subgrouping achieved a non-diploma credential
still_enrolled_cnt	How many enrolled students of the indicated subgrouping did not graduate but were still_enrolled
still_enrolled_pct	What percentage of enrolled students of the indicated subgrouping did not graduate but were still_enrolled
ged_cnt	How many enrolled students of the indicated subgrouping were awarded a "GED" diploma
ged_pct	What percentage of enrolled students of the indicated subgrouping were awarded a "GED" diploma
dropout_cnt	How many enrolled students of the indicated subgrouping discontinued their high school enrollment during the school year
dropout_pct	What percentage of enrolled students of the indicated subgrouping discontinued their high school enrollment during the school year

Your objective for Project 1 is to apply the full data science project lifecycle to the implementation of a series of regression models. Your work should include EDA, data preparation (including the use of attribute transforms as needed), feature selection, and a thorough evaluation of model performance metrics. The response variable you will be modeling is the data set's "**dropout_cnt**" attribute, which represents the number of enrolled students who discontinued their enrollment (i.e., "dropped out") from within the indicated school district | student subgroup .

Therefore, your task is to construct and compare/contrast a series of regression models (after completing the necessary EDA and data prep work) that predict the number of student "dropouts" relative to certain properties/characteristics of a given school district + associated student subgrouping. It is up to you as the data science practitioner to determine which features should be included in these models. To get started on the Project:

1) Load the provided Project1_Data.csv file to your DAV 6150 Github Repository.
2) Using a Jupyter Notebook, read the **Project1_Data.csv** 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) Construct **at least two different Poisson regression models**, **at least two different negative binomial regression models**, and **at least two multiple linear regression models**, using different explanatory variables (or the same variables if they have been transformed via different transformation methods).

At times, Poisson and negative binomial models can produce identical results. Be sure to comment on that if it happens.

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 evaluation data set and assess how well it performs on that previously unseen data.

Your first deliverable for this Project 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 (10 Points)**: Explain + present your EDA work including any conclusions you draw from your analysis regarding the integrity + usability of the data in its raw state. 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 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)**: Split the data into dedicated training and testing subsets. Explain + present your regression modeling work, including your 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 how you selected your model selection criteria. Identify your preferred model. Discuss why you’ve selected that specific model as your preferred model. Apply your preferred model to the testing subset you created during your Regression Modeling work and discuss your results. Did your preferred model perform as well as expected?
7) **Conclusions (5 Points)**

Your Jupyter 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 sub-section headers, and be free of spelling and grammar errors. Furthermore, your Pythion code should include succinct explanatory comments.

Upload your Jupyter Notebook within the provided Project 1 Assignment Canvas submission portal. Be sure to save your Notebook using the following nomenclature: **first initial_last name_Project1**" (e.g., J_Smith_Project1_). ***Small groups should identity all group members at the start of the Jupyter Notebook and each team member should submit their own copy of the team's work within Canvas.***

Your second deliverable for this Project **(10 Points)** is a short (approx. 5 minute) video presentation of your work. Your presentation should include a brief overview of your EDA findings, a high-level explanation of your data preparation + feature selection process + regression models, a summary of your model selection process, an explanation of why you chose your preferred model, and comments on the performance of your preferred model when applied to the evaluation data set.