# DAV 6150 Module 8 Assignment

## *Classification via KNN & SVM*

***** You may work in small groups of no more than three (3) people for this Assignment *****

The Module 7 Assignment made use of a data set sourced from the Federal Reserve Bank of Boston (https://vincentarelbundock.github.io/Rdatasets/doc/Ecdat/Hdma.html) containing data related to mortgage application approvals and denials. The data set was comprised of 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.

As you will recall, for that Assignment you were tasked by a large banking regulator with the development of a binary logistic regression model that can predict whether or not a given mortgage loan application is likely to be approved or denied. The regulator planned 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. For this Assignment, the regulator has asked you to revisit your work and develop an additional suite of classification models to see whether something other than a binary logistic regression model might be more effective for the designated task.

Your task for the **Module 8 Assignment** is to construct and compare/contrast a series of **K-Nearest Neighbor and Support Vector Machine 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) Ensure the **M7_Data.csv** file (provided with the Module 7 Assignment) has been loaded to your DAV 6150 Github Repository.
2) Then, using a Jupyter Notebook, read the data set from your Github repository and load it into a Pandas dataframe.
3) Perform EDA work as necessary. (If you already have a high-quality EDA from the Module 7 Assignment, you may incorporate it here. If your M7 Assignment EDA was flawed, you should repeat the EDA work and address any shortfalls identified in you M7 Assignment EDA).
4) Perform any required data preparation work, including any feature engineering or standardization adjustments you deem necessary for your work.
5) Apply your knowledge of feature selection and/or dimensionality reduction techniques to identify at least five (5) 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 two different KNN models and two different SVM 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" classification 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 (15 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) **KNN + SVM Modeling (40 Points)**: Explain + present your KNN and SVM modeling work, including your feature selection / dimensionality reduction decisions, the process by which you selected your "K" values for your KNN models and the process by which you selected your SVM models' hyperparameters and (if applicable) any kernel functions used within the SVM models. This section should include any Python code used for feature selection, dimensionality reduction, and model building.
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? How does the performance of your preferred Module 8 Assignment model compare to the performance of your preferred binary logistic regression model from the Module 7 Assignment? Be sure include any Python code used as part of your model selection work and to frame your discussion within the context of the classification performance metrics you have derived from the models.
7) **Conclusions (10 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 to your online DAV6150 GitHub directory. Be sure to save your Notebook using the following nomenclature: **first initial_last name_M8_assn**" (e.g., J_Smith_M8_assn_). Then submit the resulting web link via Canvas within the Module 8 Assignment page. ***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.***