Department of Electrical and Computer Engineering
Howard University

Washington, DC 20059

EECE 401-404 Senior Design

Spring 2022

Final Report

Submitted by:

De'Johnna Wright

Joshua Whitaker

Caelia Thomas

Instructor: Dr. Charles Kim

Advisor: Dr. Imtiaz Ahmed

Date Submitted: 04/20/2022

Summary:

Capital One tasked our team with developing a time-varying regression model to predict if a customer would be eligible for a loan depending on specified criteria. This project used Python and exclusively open source libraries to create a random forest regression model. Capital One provided the team with a datasheet consisting of data fields from a large number of customers. The relevant fields for loan eligibility were selected by the team, and the model created was trained against a datasheet. Our model was able to predict loan eligibility for customers with at most an accuracy of 90%. After verifying the functionality of the model, it was essential to test for model accuracy with drifting data. Since the data that Capital One manages may shift over time, the model may no longer perform as expected since it was only trained on the original dataset. Accounting for data drift allows for the model to retrain itself if necessary before attempting to compute loan predictions. Given the team's late start on the project and the rearranging of roles that occurred throughout the semester, the final product created was not what was originally envisioned. However, the model created was successful and the data drift computations were useful and can be used as a foundation for future iterations of this project.

Problem Statement:

The need for the Capital One team is to increase efficiency in processing data sets to determine loan eligibility and ensure that the algorithm in place will be accurate even as data changes over time.

Design Requirements:

This project depended on abiding by a set of constraints. Firstly, a machine learning approach had to be implemented. Using a machine learning algorithm will ensure that our predictions are highly accurate. Secondly, the Python libraries that are used must also be open

source libraries to avoid running into issues with copyrighted or restricted software. Also, since the project was delayed, time served as a major constraint for the team. With the initial scope of the project being complex, it was essential to determine which portions of the project to focus on in the event that the entire project could not be completed.

Aside from the project constraints, there were also external regulations on the project. The team was required to provide Capital One with monthly updates on project updates. The team also had to comply with Capital One's NDA to ensure that no sensitive information was distributed.

Solution Design:

The initial solution design was a fully functional regression model as shown in figure 1. Ideally, a user would be able to load a customer dataset into the regression, and the algorithm would take the relevant parameters and clean the data before computing the regression. The parameters identified were the annual income, employment length, and verification status of the customer. Then, the algorithm would output a visual representation of the results. The algorithm would then be tested against other datasets to improve its accuracy. Not pictured in this diagram, is the process of training the regression algorithm. This process is completed by using a training dataset, which contains the loan eligibility results, to teach the algorithm how to calculate loan eligibility. The model is then tested with a test dataset which also contains the loan eligibility results. The output of the regression is compared with the actual results in the test dataset to check the model's accuracy.

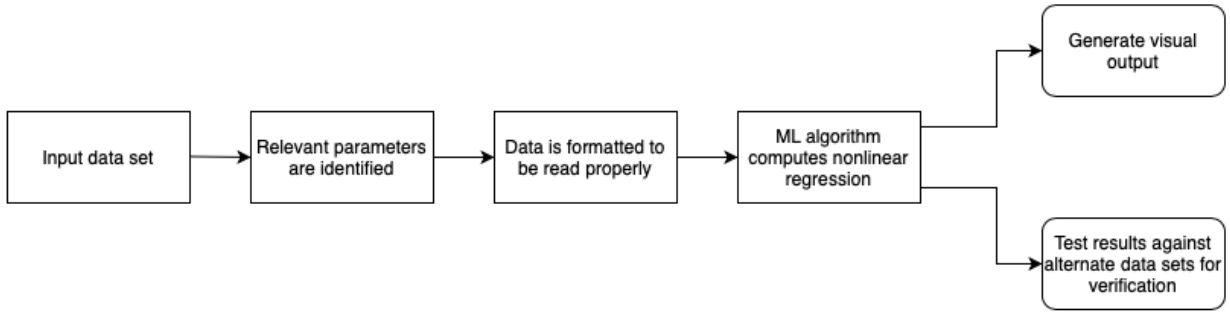

Figure 1

Unfortunately, the start of the project was delayed until January 2022. Completing a fully functional regression model that analyzed datasets of this size, was too ambitious of a goal to complete in three months. This led to the evolution of the solution design in figure 2. The refined solution design reflects the shift in focus from visualization to data drift analysis. Data drift is when values from an input dataset shift away from the values of the dataset that was used to train the regression model.

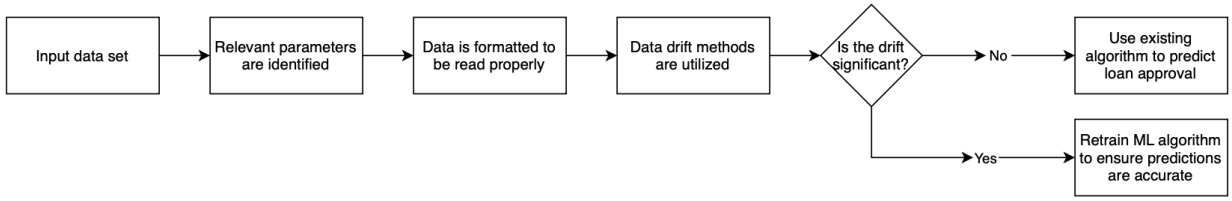

Figure 2

As before, the large customer datasets are still loaded into the algorithm, and the relevant parameters are identified and cleaned for the regression. After the data is cleaned however, a data drift detection method is utilized to determine if the drift is significant enough to affect the model's accuracy. The model is still trained on the original dataset before this process begins. If the data drift detection method finds that the drift is not significant, then the algorithm can

continue to predict the loan eligibility. If the drift is significant, then the algorithm has to be retrained to handle this new dataset.

Agile Plan:

The original plan for the spring 2022 semester is detailed in figure 3. When the original end-deliverable was a Time-Varying Regression Model, the semester was broken down into four sprints of work. The first sprint would result in a prototype regression algorithm, the second would identify the best data drift detection method for the model itself. The last two sprints would be centered around fine-tuning the accuracy of the model, and generating a visualization of the loan eligibility. This would be accomplished by creating a feedback loop with the data drift detection and accuracy of the model. If the drift is affecting the model's accuracy, then the feedback loop would prompt the algorithm to be re-trained in order to maintain a high accuracy level.

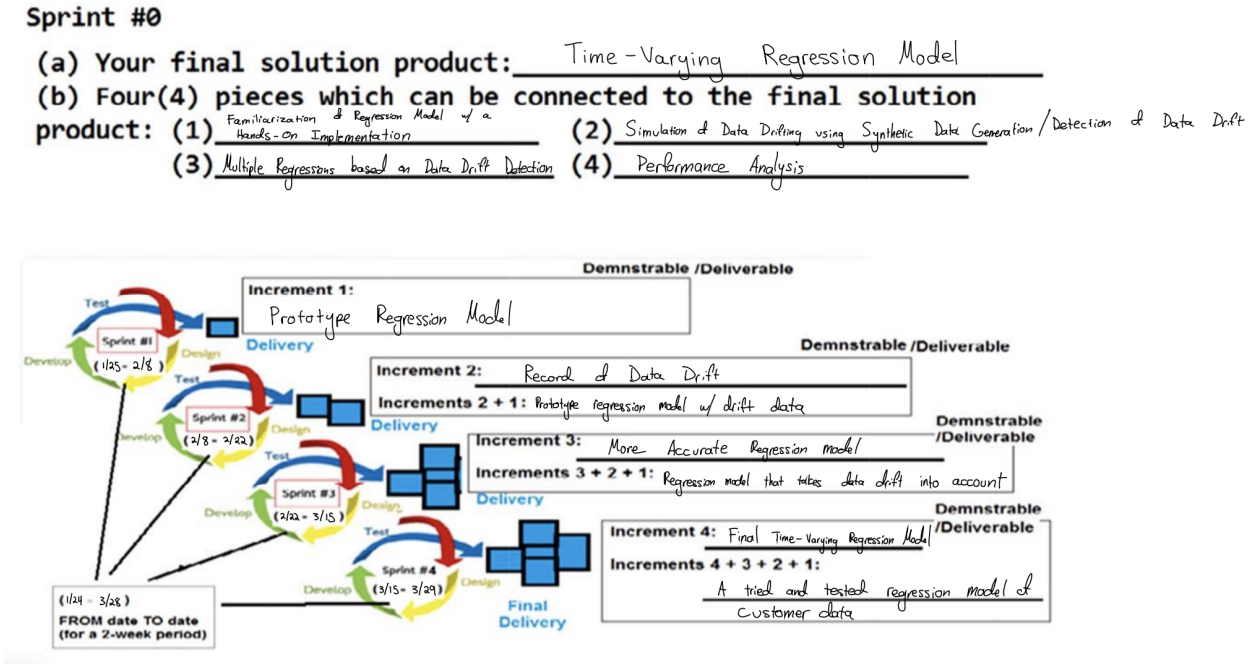

Figure 3

Project Implementation Process:

The first sprint focused on determining loan eligibility for Capital One account holders. This was completed by initially removing excess characters and converting non numerical statements to numbers. As seen below in figure 4 for verification status account holders were either not verified, source verified or verified and these statements were assigned to either 1,2 or 3.

verification_status	verification_status
Not Verified	1
Not Verified	1
Source Verified	2
Source Verified	2
Source Verified	2
Not Verified	1
Not Verified	1
Source Verified	2
Verified	3

Figure 4

From here we used Python libraries to create random forest regression algorithms which focused only on employment length, annual income and verification status to determine whether a customer would be approved for a loan or not. After testing the model using the test datasheet against the training datasheet our values were determined to be at most 90% accurate using a small sample size. While going through the process of determining the accuracy of the regression algorithm we altered our code using different threshold values. As seen in figure 5, the higher the threshold value the more confident the regression was which resulted in a less accurate model.

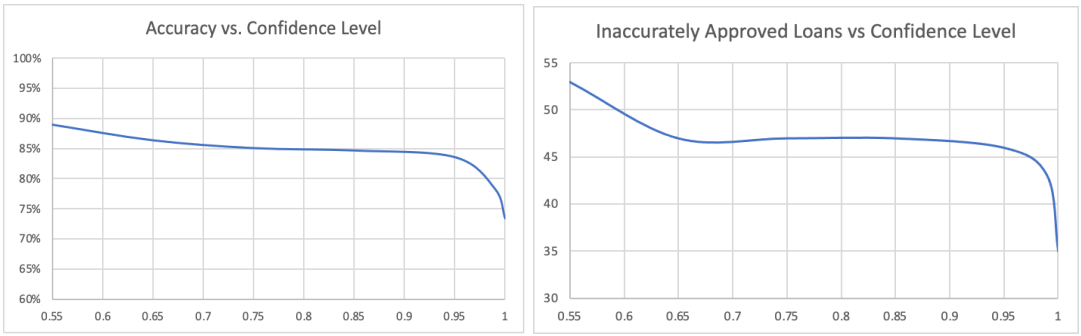

Figure 5

The second sprint directed the team's efforts to computing data drift. A total of five different data drift methods were tested: ADWIN, Population Stability Index, K-S, Page Hinkley, HDDM_W. Initially, the training dataset that the data drift methods were tested on had the annual income values' mean shifted by 14000. The ADWIN, Page Hinkley, and HDDM_W methods all detected data drift by flagging values which exceeded a predetermined threshold. The three methods differ in how they flag values, so the amount of values each method detects differ greatly from the others. The K-S and PSI methods determine how similar the training dataset and original dataset fields are. The K-S method will either reject data fields that differ too greatly, or accept data fields that are mostly similar. The PSI method instead returns a value determining how similar the fields are. A PSI greater than 0.2 indicates a data field which has changed significantly, a PSI between 0.2 and 0.1 indicates a moderate shift, and a PSI less than 0.1 indicates little to no shift. Below, Table 1 and Table 2 show the values obtained from the data drift testing, and Table 2 displays a histogram comparing the original dataset to the modified dataset.

Method	Results
PSI	Stability index: 0.22
K-S	K-S value: 0

Table 1

Method	Results
Adwin	7 changes detected
Page Hinkley	2714 changes detected
HDDM_W	18420 changes detected

Table 2

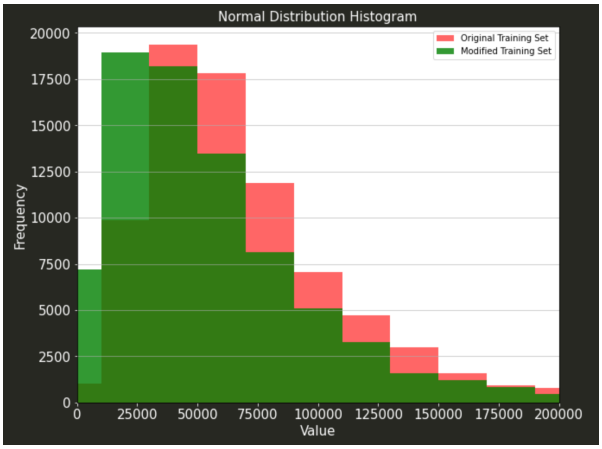

Figure 6

The third sprint continued with the data drifting techniques being applied to the three modified data sets to analyze and take note on how the data drifts techniques operate on different datasets. In this stage we determined with our advisor that PSI was the most effective data drifting method since it did indeed output the most efficient information whereas the other 3 methods output an abundance of varying information that was not useful. Figure 7.1 displays a histogram for dataset 2 with a PSI value of 1.1, figure 7.2 is a histogram for dataset 3 with a PSI

value of 1.77, and figure 7.3 is a histogram for dataset 4 with a PSI value of 5.25. The large PSI values indicate there was a significant shift in the data.

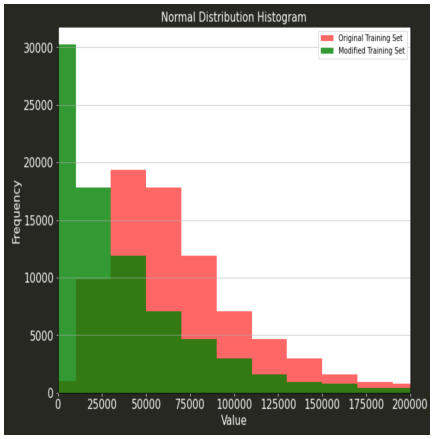

Figure 7.1

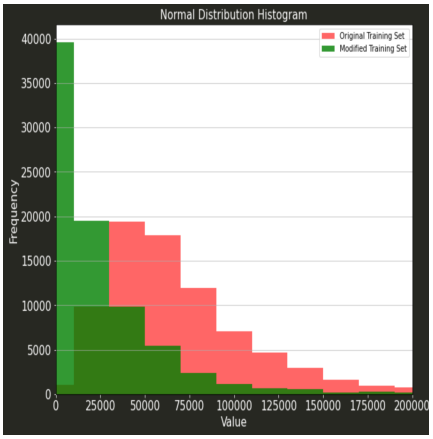

Figure 7.2

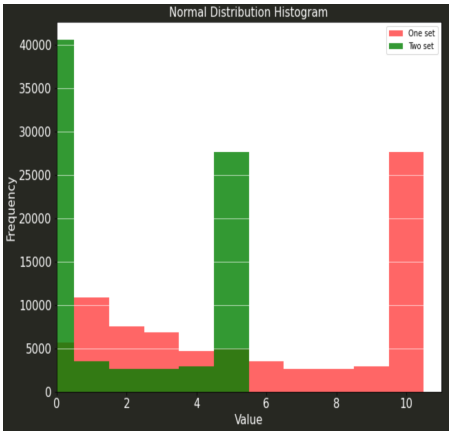

Figure 7.3

During the fourth sprint there was some reconstruction of the project, the team was to focus solely on data drifting research in addition to concept drift. The original algorithm was run against the 3 modified datasets to evaluate how the drifted data affects the random forest prediction.

Dataset	**PSI**	**Accuracy (%)**	**Inaccurately Approved Loans**
1	0.22	82.1	45
2	1.1	69.89	31
3	1.81	66.7	28

Table 3

Conclusions:

The original goal of the Capital One project was to create a fully functional Time-Varying Regression Model. This was a goal that could not be completed in three months, but a lot of progress was made that can aid future VIP project groups. The prototype regression algorithm was created, and the data cleaning process and parameter selection was finalized to create an initial loan eligibility determination. Different data drift detection methods were tested in compatibility with the random forest regression that was used, and PSI was identified as the most relevant and compatible detection method.

Concept drift research is the current stage of the project at the time of writing this report, and the accuracy performance of the current model prototype is a fantastic springboard for the VIP project group in fall 2022.

References:

AlindGupta@AlindGupta. "Python: Linear Regression Using Sklearn." GeeksforGeeks, 26 Jan. 2019, https://www.geeksforgeeks.org/python-linear-regression-using-sklearn/.

"Data Drift Detection: Importance of Data Drift Detection." *Analytics Vidhya*, 15 Oct. 2021, https://www.analyticsvidhya.com/blog/2021/10/mlops-and-the-importance-of-data-drift-detection/.

"Data Drift and Machine Learning Model Sustainability." *Data Drift and Machine Learning Model Sustainability* |, 22 Feb. 2021, https://www.analyticsinsight.net/data-drift-and-machine-learning-model-sustainability/.