Once the Lending Club data for all loans from 2007 to Q3 of 2019 has been collected and cleaned, the data set is ready for the application of machine learning. Several different machine learning algorithms were utilized to explore and make predictions on the dataset. Supervised Learning algorithms such as Lasso Regression, Random Forest, and SVM were used to identify key features of the dataset and predict loan status, while unsupervised machine algorithms such as KMeans clustering and Primary Component Analysis were used to explore the data points.

Because the cleaned and wrangled dataset contains 51 features, Lasso Regression would play a vital role in determining which features have the most predictive power when considering loan status. The Lasso regression revealed that the amount of outstanding principal (out_prncp) and the last payment amount (last_pymnt_amnt) were the features with the largest direct correlation to a loan entering default. On the other hand, loan amount (loan_amnt), total payment (total_pymnt), and total interest received (total_rec_int) were the features with the most negative correlation to loans entering default. Swarm plots for these features show the distribution of the respective features for loans not in default overlayed with the respective features of loans in default.

To further explore the data set and look for trends in the individual data points, KMeans clustering was applied to the data set. In order to determine the best number of clusters for the dataset, the sum of squares was computed for different values of n_clusters. Using the knee/elbow method revealed that 3 or 4 clusters significantly reduced the squared sum of distances from each point to their respective cluster centers. Because the elbow in the graph was not sharp enough to exactly pinpoint the best number of clusters for KMeans, silhouette analysis was also applied to the dataset. In combination with the Primary Component Analysis that reduced the dimensionality of the dataset to two features, the scatter plot of the reduced dataset showed four sets of clustered data points.

To make initial predictions on the default status of loans, a Random Forest was trained with its default hyperparameters. When evaluating the model, the confusion matrix and specifically the number of True Positives revealed the model's lack of ability to correctly predict if a loan will default. More specifically, the number of True Positives and False Positives were both zero, indicating that the model predicted that all loans will not default. After further investigation, the cause of the problems is the presence of a severe class imbalance in the target column of the data set.

A sample of the dataset revealed that 619,995 loans did not default and only 853 loans did default. This is understandable as Lending Club only wants to issue loans that will be repaid in-full and avoid default, therefore minimizing the number of loans that

enter default. To fix the issue of class imbalance, SMOTE from imbalanced-learn was applied to oversample the number of default loans. Doing so will decrease the overall accuracy of the model, but will increase the model's ability to correctly predict if a loan will default. From a business perspective, it is not a major issue if loans that will not default are classified as a loan that will default, but it would be very detrimental to the lending institution if loans that will default go unrecognized and are classified as a loan that will be successfully repaid. Therefore, it is in the best interest of the lending institution to maximize the model's ability to correctly identify the loans that will default.

After implementing the Synthetic Minority Oversampling Technique to the training data, the newly trained model performed with 71% accuracy. But more importantly, the number of True Positives increased and the number of False Positives decreased in the confusion matrix. The area under the ROC curve was calculated to be 0.7, which indicates the model has a significant level of predictive power.

To practice building a supervised classifier and tuning its hyperparameters, a Support Vector Machine was also implemented to predict if a loan will or will not default. To maximize the number of True Positives that the SVM is able to predict, several modifications had to be made to the data and the hyperparameters.

Unlike a random forest, a support vector machine requires the features to be scaled in order to avoid distance bias within the features. The features of the dataset were scaled with scikit-learn's StandardScaler, and the initial model was trained on the data. Again, due to class imbalance of the target variable, SMOTE was used to oversample the data for loans that did default. This improved the model because the model no longer predicted that all loans will not default.

To further improve the model, Grid Search Cross Validation was implemented to identify the best hyperparameters for the SVM. After individually training models on the scaled data with all of the possible combinations of *C* and *gamma*, the GridSearcCV returned the best combination of hyperparameters.

Due to the size of the Lending club dataset, the SVM was trained on a sample of the original data set. A pipeline was created to scale the data, identify the best parameters with grid search and make predictions on the data with an SVM trained with the best combination of hyperparameters. With the pipeline running successfully, the complete data set can be fed in the data to improve the quality of the model and its predictive power.