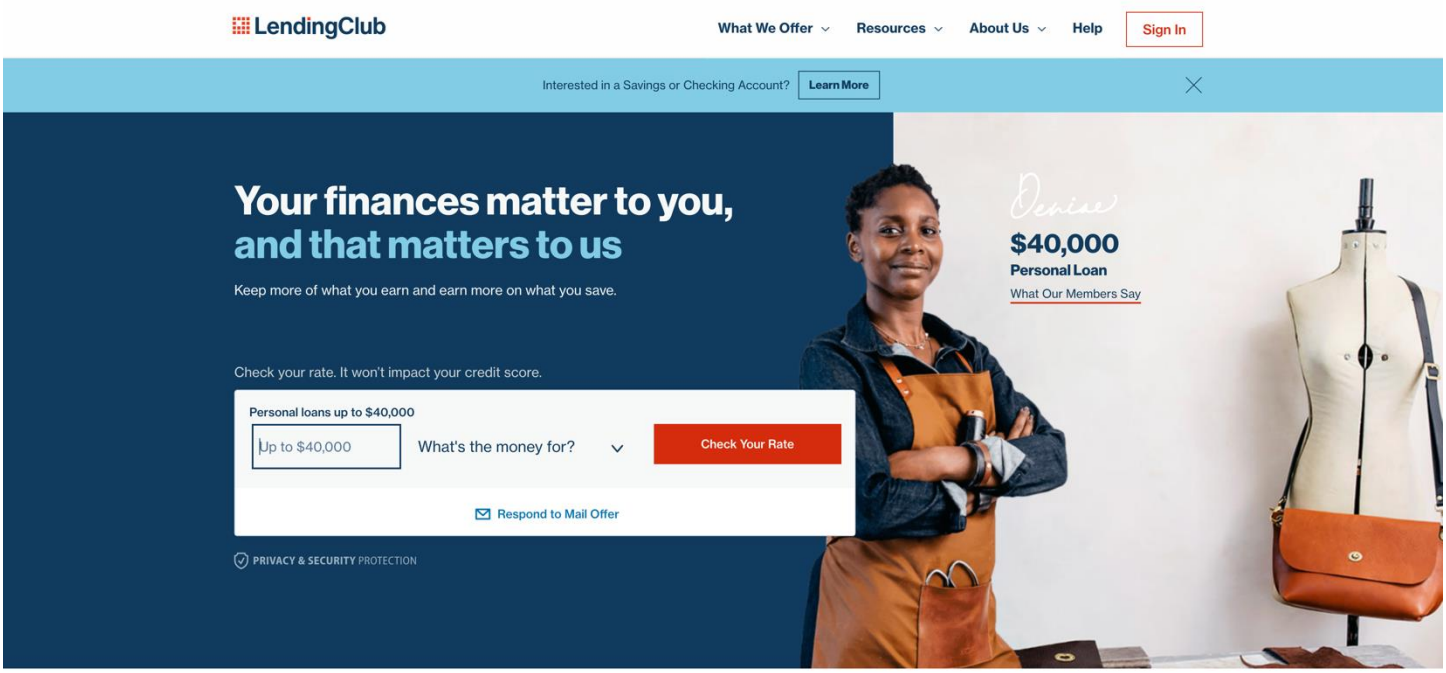
HW Assignment 2 (A2): Classification Models
CS6140: Machine Learning Fall 2024
Due Date: Thursday, _____
(__ points)



Scenario

LendingClub is an online peer-to-peer lending platform that connects borrowers with individual investors, acting as an intermediary. It was founded in 2006 and is one of the pioneers in the peer-to-peer lending industry. LendingClub offers various types of loans, including personal loans, business loans, and auto refinancing, among others. Borrowers can apply for loans by filling out a profile, after which LendingClub assesses their credit risk and assigns an interest rate to the loan. Individual or institutional investors can then browse the platform for loans that meet their investment criteria and choose to fund those loans. Once a loan is funded, the borrower receives the money and begins making monthly payments, which are distributed back to the investors.

In 2015, LendingClub faced a major controversy when it was discovered that the company had altered loan data, leading to the resignation of the CEO and increased regulatory scrutiny. This event raised concerns about the integrity of loan data, which is crucial for building reliable credit risk models.

You are provided with datasets containing approved loans from 2014 to 2016. You are also provided with the combined dataset and a data dictionary for their analysis. This should give them everything they need to get started on their assignments.

Your task is to build classification models to compare model results and signals between these two periods, exploring how the controversy might have impacted loan grading.

Task

Classification is a fundamental task in supervised learning, where the goal is to assign predefined labels to new data based on patterns learned from a labeled dataset. In this assignment, you will build and evaluate different classification models to predict the grade of loans based on their features. You will analyze how the classification performance may differ between two datasets from before and after LendingClub's 2015 controversy, which involved data integrity issues that could have impacted loan grading practices.

Task

Classification is an essential operation of supervised learning prediction problems. Classification refers to assigning predefined labels or categories to new, unseen data based on the patterns learned from a training dataset. It is a supervised learning approach, meaning the model is trained on a labeled dataset. The goal is to map input features to discrete output labels. Classification tasks are often contrasted with regression tasks, which aim to predict continuous numerical values rather than discrete labels. There are numerous algorithms that accomplish those tasks. You will explore several foundational algorithms in this assignment, including Logistic Regression, KNN, and SGDClassifier. Logistic regression predicts a binary outcome (e.g., yes/no, pass/fail, true/false) and occasionally used for multi-class classification, based on a set of independent variables.

K-Nearest Neighbors (KNN) is a simple, non-parametric algorithm. In the KNN algorithm, a data point is classified based on how its neighbors are classified. Neighbors are chosen based on a distance metric. The SGDClassifier stands for Stochastic Gradient Descent Classifier and is commonly used for large-scale and sparse machine learning problems. It is part of the scikit-learn library in Python.

The above classification algorithms are powerful tools used for a variety of applications such as predicting customer churn, credit risk, medical diagnosis, spam detection, text classification and NLP. You are given two-quarters of data 2018 to build classification models. The data dictionary is also provided to you. The primary objective is to explore the above classification techniques on two types of responses: **High-Low** grade of loan and **High-Medium-Low** grades of loan.

High-Low schema:

High: Grade = A or B
Low: Grade = D or E or F or G

High-Medium-Low schema:

High: Grade = A or B
Medium: Grade = C
Low: D or E or F or G

Perform the following classification tasks and submit a fully executed document providing solutions to the following queries. Make sure to add commentary and analysis for each step.

1) **Summary:**
   Perform EDA and share insights you learned focusing on any differences between the two time periods of concern. This is an important step that might guide you in creating the final model.

2) **Preprocessing**:
   a. Standardize the numerical features and encode the categorical features
   b. Identify and remove up to 1% of rows as outliers based on standardized `dti`, `annual income`, and `delinq_2yrs` variables
      Hint: Standardize the columns, add them up, and then identify outliers using IQR method.

3) Classification Task:
   a. Split the data into Train-Validate-Test
   b. Build a logistic model to accurately predict the High-Low response.
      i. Discuss which variables are significant and how that may help Lending Club make predictions.
   c. Build two KNN models predicting both responses (High-Low and High-Medium-Low)
   d. Build two SGDClassifier models predicting both responses (High-Low and High-Medium-Low)
   e. Compare the accuracy of all final models (overall accuracy, precision, recall). Discuss which model you would recommend.

Expected Output
Please submit a fully executed Jupyter notebook clearly identifying question number and steps. Make sure to add proper commentary to your solution.