SEA600- Assignment 1

Total Mark:	100 marks
Submission file(s): (do **NOT** zip files)	Milestone submissions: - Code.ipynb - Assignment1.docx with draft responses to questions in “Final submission” section Final submission: - Code.ipynb - Report.docx (or pdf) - Group.docx

Please work in **groups** to complete this lab. This assignment is worth 15% of the total course grade and will be evaluated through your written submission, as well as the lab demo. During the lab demo, group members are *randomly* selected to explain the submitted solution. Group members absent during the lab demo will lose the demo mark.

Please submit the submission file(s) through Blackboard. Only one person must submit for the group and only the last submission will be marked.

Contents

Description

In this assignment, you will design a solution using Machine Learning techniques to predict target values for regression and binary classification problems. You will compare simple models with advanced models based on performance measures as well as resource utilization. To improve the performance further, you will apply feature engineering and hyperparameter tuning. Furthermore, you will analyze your results to pick the most suitable model for deployment (test).

Please review the questions in the “Final submission” section to guide you in each milestone. Create a draft of your responses each week. Use these drafts to create a clean technical report for the final submission.

Milestone I: Problem Definition and Baseline Methods

1. Review chapter 2 of HOML book:

   https://libaccess.senecacollege.ca/login?url=https://learning.oreilly.com/library/view/hands-on-machine-learning/9781098125967/ch02.html/?ar&email=^u

   and the code:
   https://colab.research.google.com/github/ageron/handson-ml3/blob/main/02_end_to_end_machine_learning_project.ipynb

2. Review the Machine Learning project checklist (appendix A of HOML book):
   https://libaccess.senecacollege.ca/login?url=https://learning.oreilly.com/library/view/hands-on-machine-learning/9781098125967/app01.html/?ar&email=^u

3. Choose **one regression OR one binary classification problem** and suitable datasets. Each group must have unique datasets to work on and get approval, therefore coordinate with your professor ASAP.

   In addition to addressing an interesting ML problem, you want to find a dataset that has enough samples for data splits and efficient training. For some datasets, you may need more time for preprocessing and cleaning.

   Resources (more in HOML- chapter 2):

   - **Kaggle**: Kaggle hosts a wide variety of datasets contributed by the community: www.kaggle.com/datasets.
   - **UCI Machine Learning Repository**: The UCI Machine Learning Repository is a collection of datasets maintained by the University of California, Irvine: https://archive.ics.uci.edu/ml/index.php or https://archive-beta.ics.uci.edu/.

- **Google Dataset Search**: Google Dataset Search is a search engine specifically designed to help researchers and data scientists find datasets: https://datasetsearch.research.google.com/ .
- **OpenML**: OpenML is an open-source platform that hosts a large collection of datasets and machine learning tasks: www.openml.org.
- **Data.gov**: Data.gov is a U.S. government initiative that provides access to various datasets from federal agencies covering a range of topics, such as health, climate, finance: www.data.gov.
- **Statistics Canada:** Economic, social and other Canadian data is available here: https://www150.statcan.gc.ca/n1/en/type/data?MM=1
- **GitHub**: Although GitHub is mainly designed for version control and collaborative development, some members host datasets on its platform as well. You can search for datasets using the search option or topic tags: https://github.com/.
- **Makeover Monday Datasets:** You can find many datasets on https://www.makeovermonday.co.uk/data/ which is a community for collaboration on data visualization.

(ANSWER):
For this project we have used the dataset modified lending club available here https://figshare.com/articles/dataset/Lending_Club/22121477?file=39316160
This dataset has two hundred thirty-six thousand eight hundred forty-seven records in its training set and ninety five thousand and twenty in its training dataset. The dataset is contains Lending Club's accepted and rejected loan data from 2007 through 2018.
The features in the dataset represent various aspects of loan applications, such as the date issued (`issue_d`), applicant's credit subgrade (`sub_grade`), loan term (`term`), home ownership status (`home_ownership`), FICO score ranges (`fico_range_low`, `fico_range_high`), total credit lines (`total_acc`), public records (`pub_rec`), revolving line utilization rate (`revol_util`), annual income (`annual_inc`), interest rate (`int_rate`), debt-to-income ratio (`dti`), loan purpose (`purpose`), mortgage accounts (`mort_acc`), loan amount (`loan_amnt`), application type (`application_type`), installment amount (`installment`), verification status (`verification_status`), bankruptcies (`pub_rec_bankruptcies`), state (`addr_state`), initial listing status (`initial_list_status`), revolving balance (`revol_bal`), unique ID (`id`), open credit lines (`open_acc`), employment length (`emp_length`), loan status (`loan_status`), and time to earliest credit line (`time_to_earliest_cr_line`). These features can be used to assess loan approval likelihood and borrower risk profile.
What we are trying to do is make a binary classifier to predict mortgage approval based on loan status and home_ownership

4. Split the data and set aside the test set for the final step (step 10). You mut not use the test set in any step prior to step 10.

   data is already splitted

5. Train **two simple classifiers** (for the binary classification problem) or **two simple regressors** (for the regression problem) without tuning or optimizing the hyperparameters. These will be your baseline methods.
   Hint: You can refer to chapters 3, 4, or 8 of MLWP book.

6. Evaluate the performance of above trained models.
   Hint: You can refer to chapters 5 to 7 of MLWP book.

7. Evaluate the resource utilization of above trained models.
   Hint: You can refer to chapters 3 and 4 of MLWP book.

Milestone II: Advanced Models

8. Based on what you have learned in Milestone I, implement up to **two additional classifiers** (for the binary classification problem) or **up to two additional regressors** (for the regression problem).
   Hint: You can refer to chapters 8-9 of MLWP book, or chapters 2-5 of HOML book.

9. Evaluate the performance and resource utilization of above trained models.

Milestone III: Feature and Model Engineering

10. Investigate the features used in your models. Is feature engineering needed? Is manual feature engineering possible and/or useful? How about automatic methods? If yes, show their effects and results. If no, explain why.
    Hint: You can refer to chapters 10 and 13 of MLWP book, or chapters 2-5 of HOML book.

11. Pick the best classification and the best regression model trained so far. Only for these models, implement and show tuning of hyperparameters.
    Hint: You can refer to chapter 11 of MLWP book, or chapters 2-5 of HOML book.

12. Compare the performance of your tuned modes with the models in milestone I and II on the test set.

Final Submission

For your final submission, please finalize you code file (Jupyter notebook) and create a technical report.

- In your code file, clearly mark the code for each milestone and question, and include comments with sufficient detail.

□ In your report, include details regarding questions in Milestone I to III. Include responses to questions from the machine learning project checklist. Please note that this should be in a technical report format, not Q&A.

  - Please see these links for some hints:
    https://www.peo.on.ca/sites/default/files/2019-11/Engineering-Report-Guide.pdf
    https://www.theiet.org/media/5182/technical-report-writing.pdf

□ Your report must at a minimum include explanations about the following details:

- Problem & data description (5%)
    - What is the objective?
    - How does it translate into a ML problem?
    - What are the implementation constraints (e.g. resource utilization)? How does each constraint affect your design decisions?
    - Are there any societal, economic, health and safety, or regulatory factors that could affect your design decisions?
    - Dataset description: Include details such as number of samples, number of features, web address of dataset, etc.

Problem & Data Description

Objective: The project aims to predict whether a loan will be approved using historical lending data from LendingClub, a peer-to-peer lending company. This is a binary classification problem where the outcome is whether a loan is approved or not.

ML Problem Translation: The objective translates into a supervised machine learning problem where the goal is to classify loan applications as approved or not approved based on features extracted from historical data.

Implementation Constraints:

Resource Utilization: Models need to be resource-efficient due to constraints on computation time and memory. This affects the choice of algorithms, favoring those with lower complexity.

Societal Impact: The model's predictions could impact individuals' financial opportunities, necessitating high accuracy and fairness in predictions to avoid discriminatory outcomes.

Regulatory Compliance: Adherence to financial regulations such as the Equal Credit Opportunity Act (ECOA) is necessary to ensure non-discriminatory lending practices.

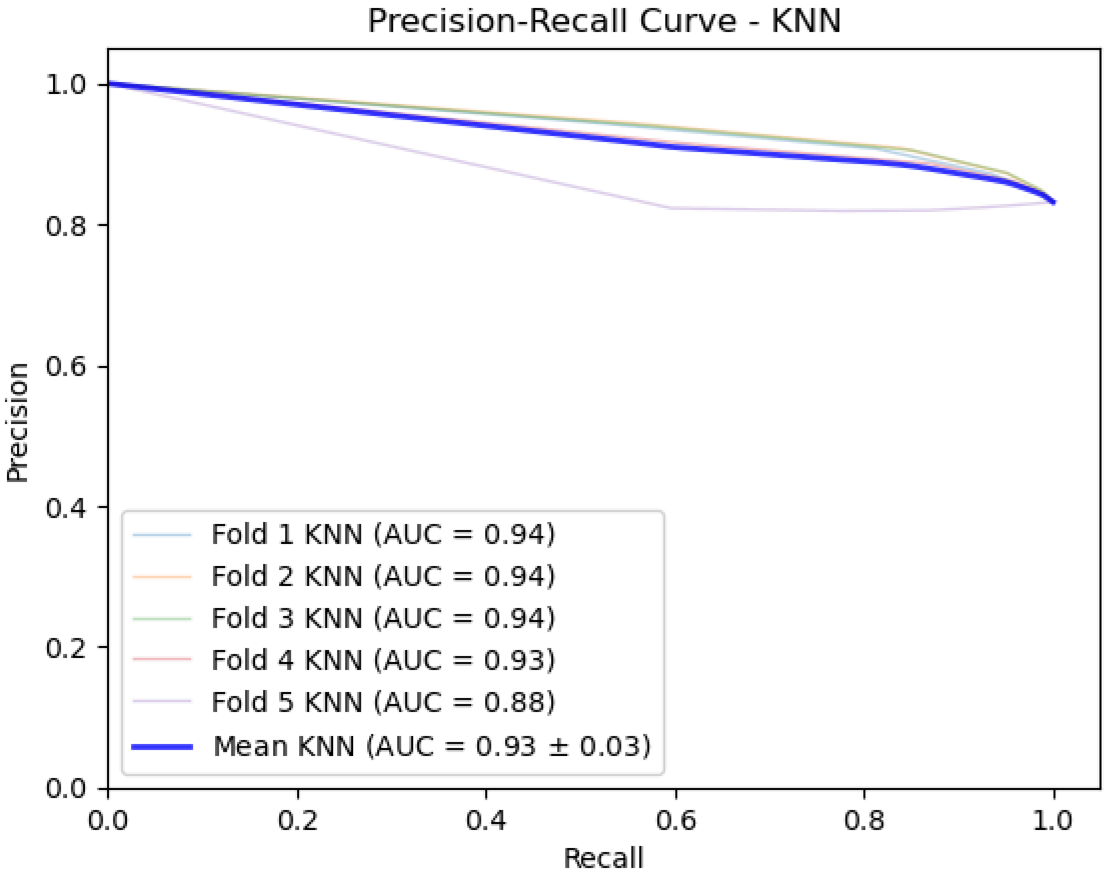
Dataset Description: The dataset contains loan application records with various features, including personal financial information and loan details. The number of samples and features, as well as the source of the dataset, will be detailed in the report.

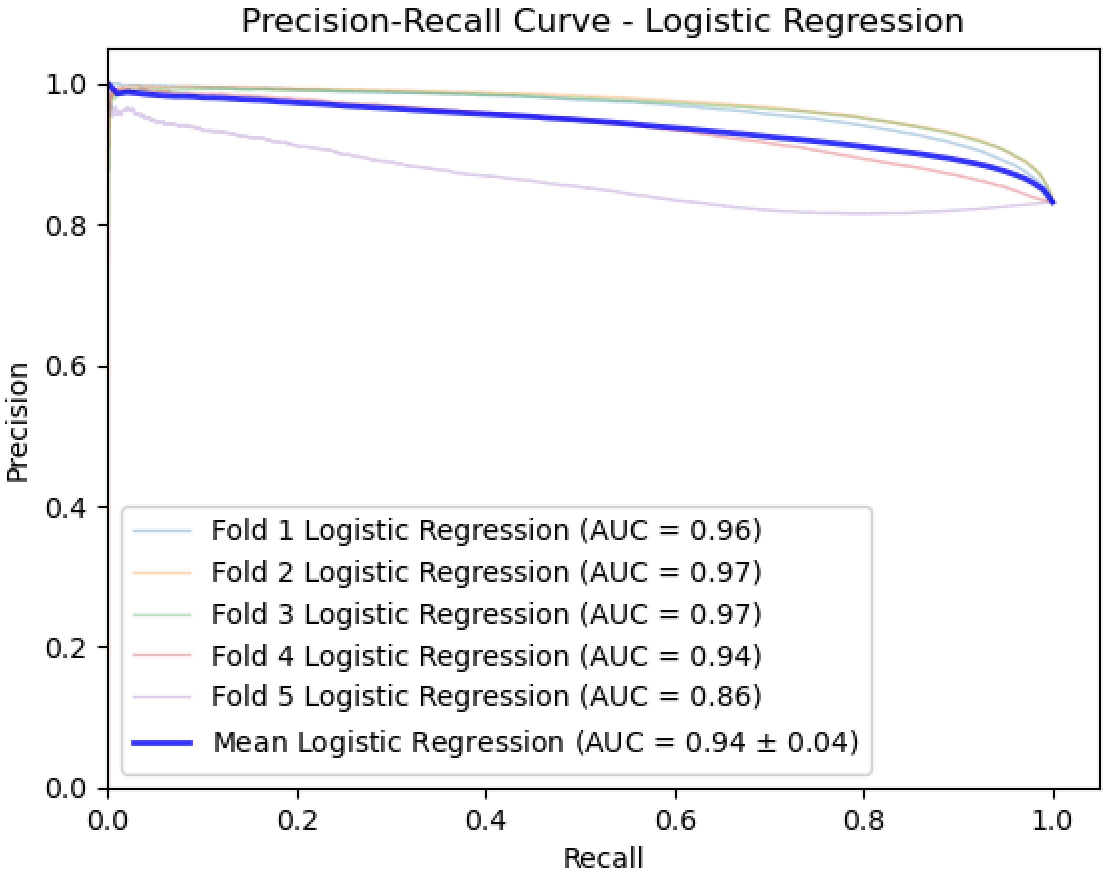
- Data exploration and preparation (5%)
    - What are the features and target variables?
    - What do the data visualizations tell you?
    - Did you need to clean the data? Why and how? What were your options? Why did you choose your method(s)?
    - How did you split the data? What were your options? Why did you choose your method(s)?
- Baseline methods, evaluations and analysis of the results (milestone I) (10%)
    - What methods did you use as baseline and why?
    - What did you conclude?

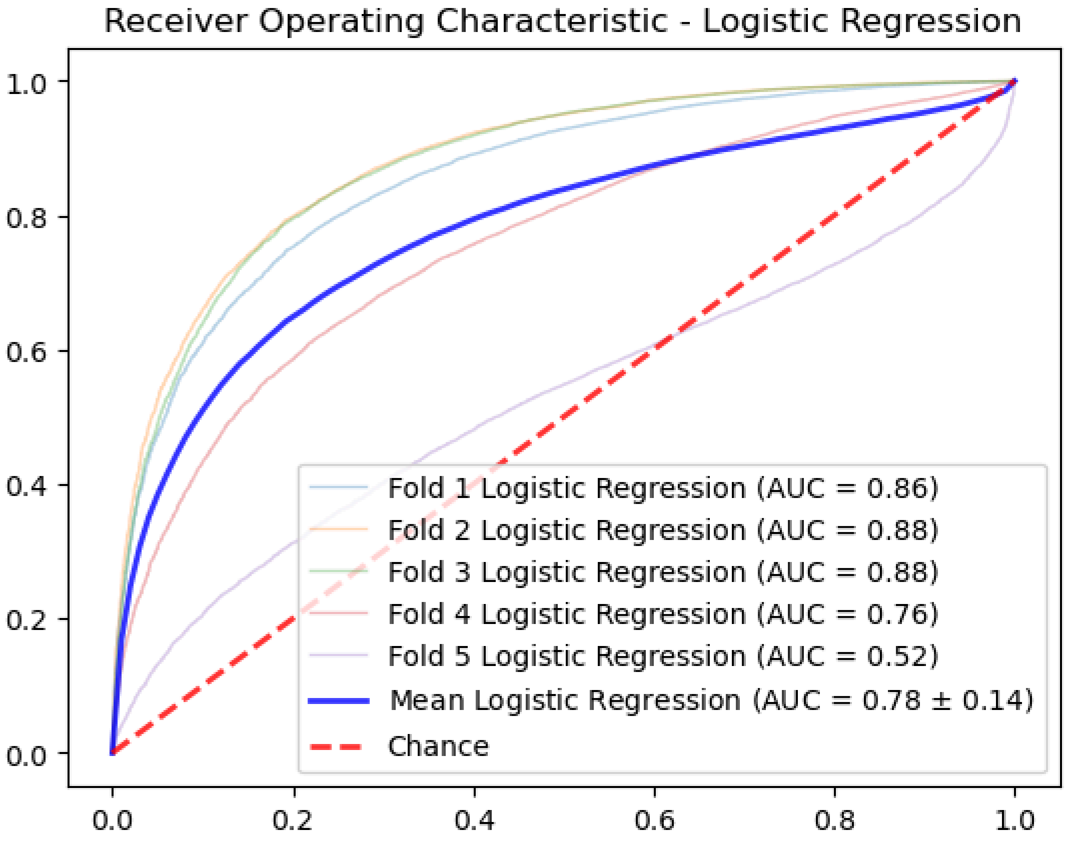
Baseline Methods: Simple models such as logistic regression and KNN served as baselines. These models are quick to implement and provide a benchmark for performance.

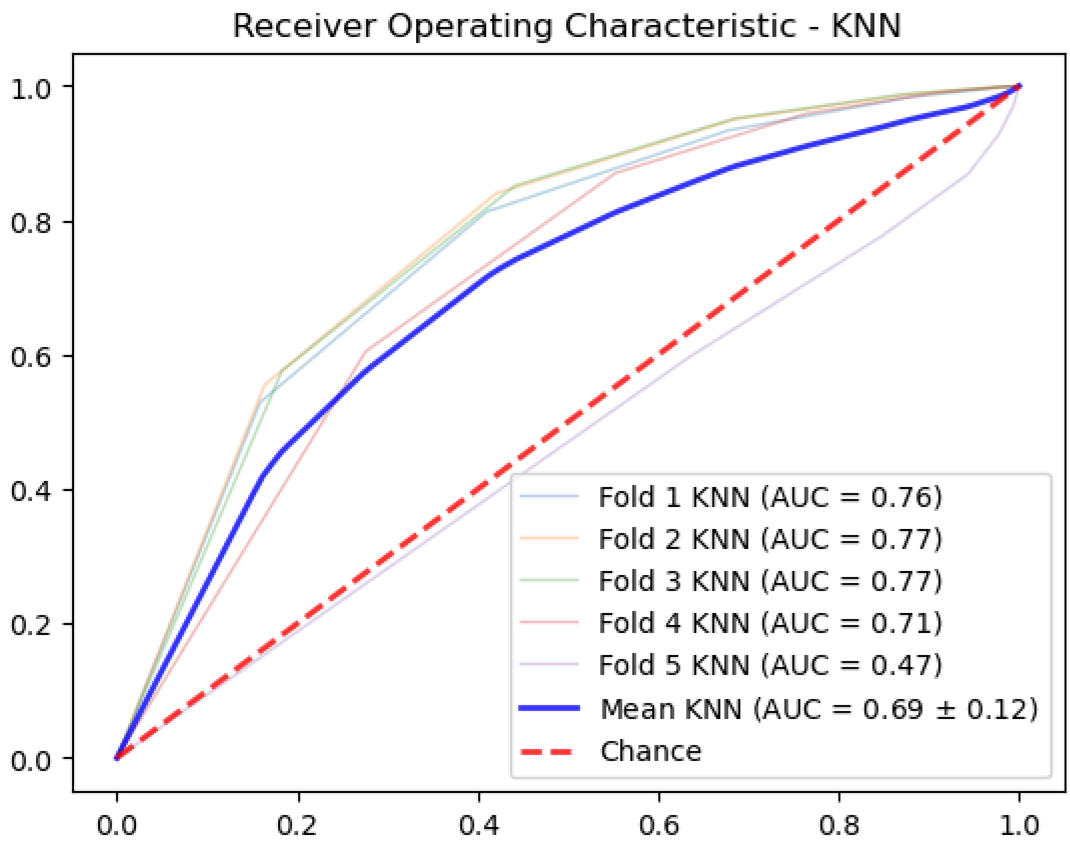
```
Time taken for training KNN Classifier: 0.08687520027160645 seconds
peak memory: 763.15 MiB, increment: 142.79 MiB
Time taken for training Logistic Classifier: 1.1194474697113037 seconds
peak memory: 768.02 MiB, increment: 147.64 MiB
```

Conclusion: They provide similar accuracy, precision and have the same memory usage


Precision-Recall Curve - KNN
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Fold 1 KNN (AUC = 0.94)
Fold 2 KNN (AUC = 0.94)
Fold 3 KNN (AUC = 0.94)
Fold 4 KNN (AUC = 0.93)
Fold 5 KNN (AUC = 0.88)
Mean KNN (AUC = 0.93 ± 0.03)

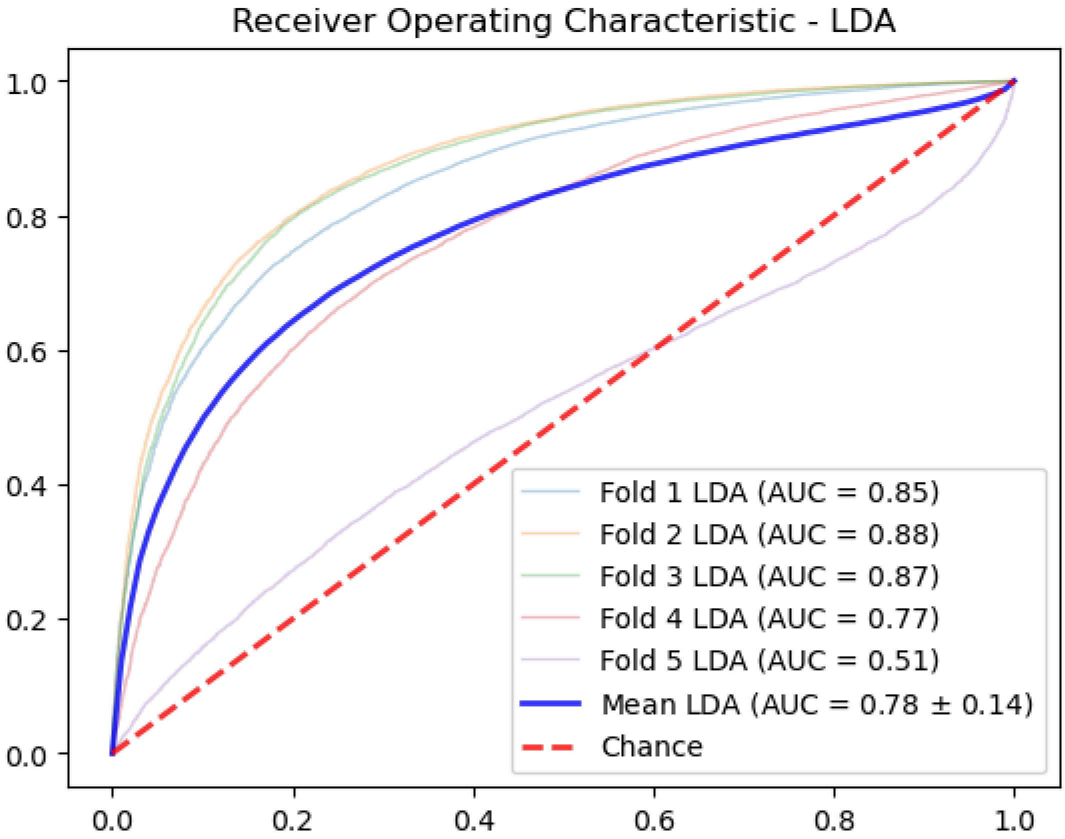
Precision-Recall Curve - Logistic Regression
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Fold 1 Logistic Regression (AUC = 0.96)
Fold 2 Logistic Regression (AUC = 0.97)
Fold 3 Logistic Regression (AUC = 0.97)
Fold 4 Logistic Regression (AUC = 0.94)
Fold 5 Logistic Regression (AUC = 0.86)
Mean Logistic Regression (AUC = 0.94 ± 0.04)

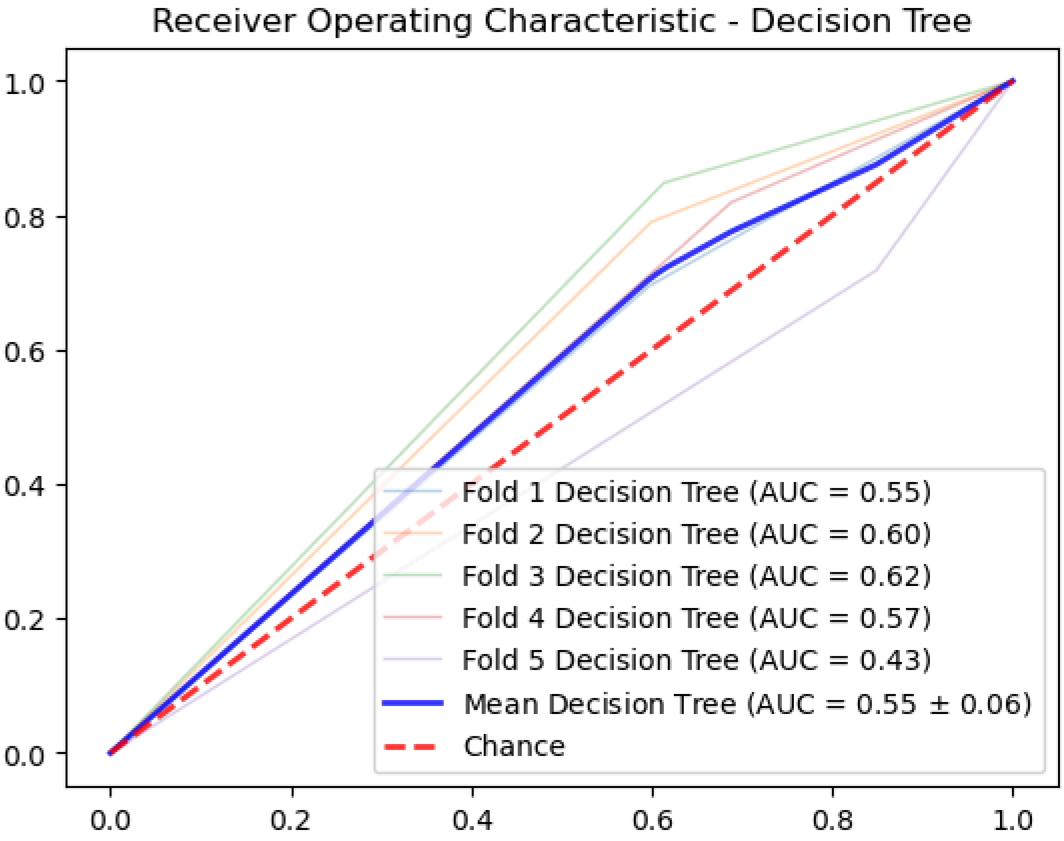

Receiver Operating Characteristic - Logistic Regression
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Fold 1 Logistic Regression (AUC = 0.86)
Fold 2 Logistic Regression (AUC = 0.88)
Fold 3 Logistic Regression (AUC = 0.88)
Fold 4 Logistic Regression (AUC = 0.76)
Fold 5 Logistic Regression (AUC = 0.52)
Mean Logistic Regression (AUC = 0.78 ± 0.14)
Chance

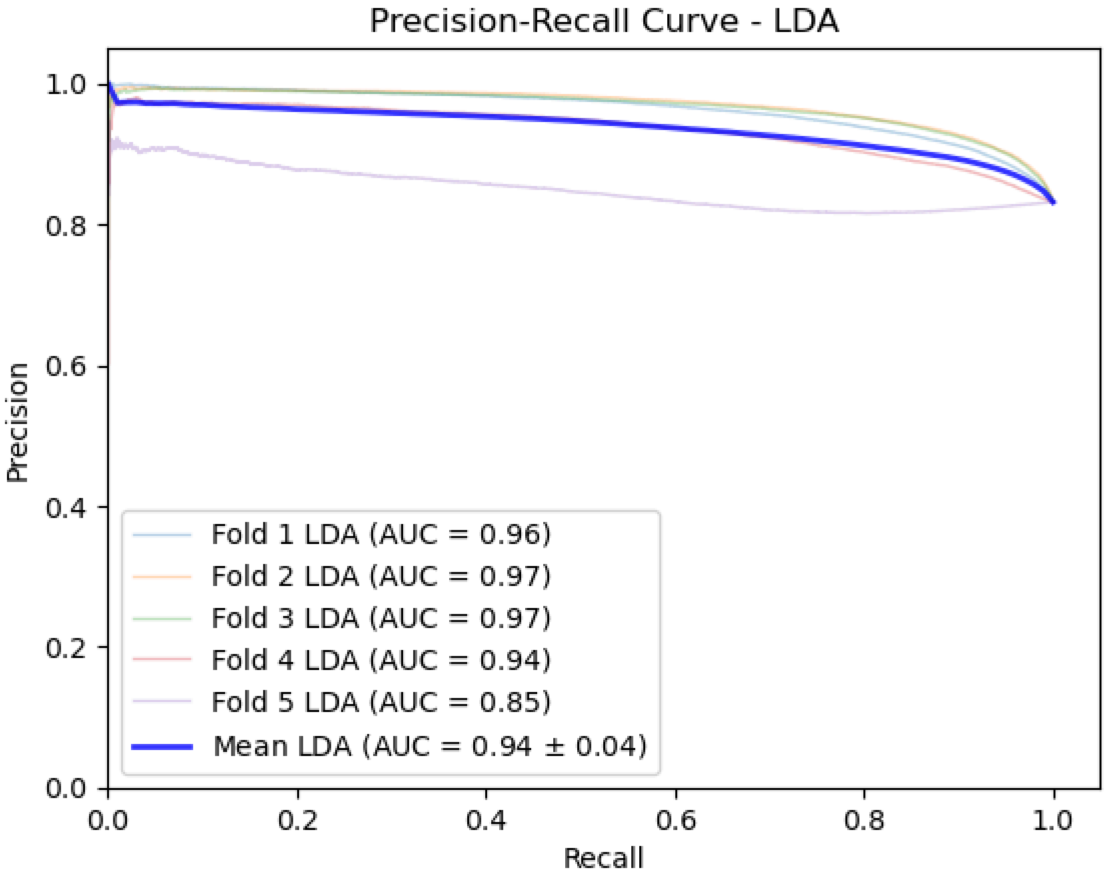


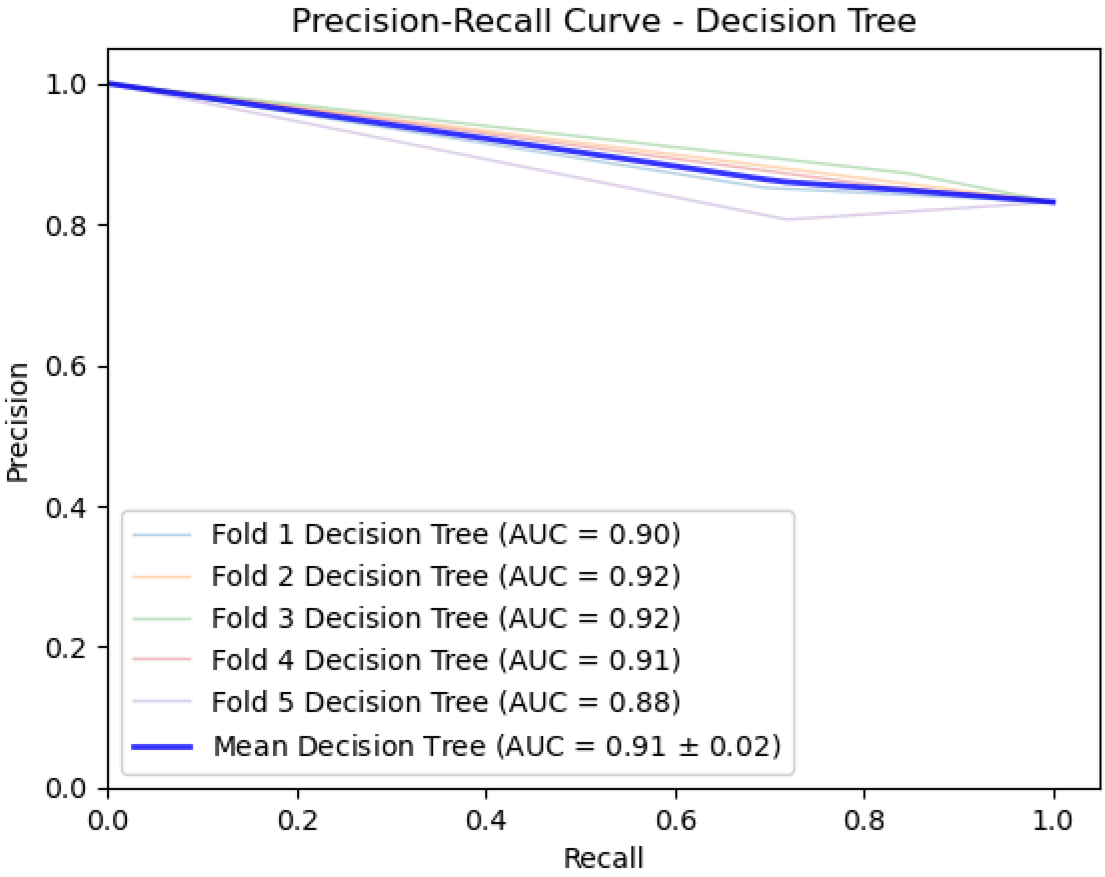
- <u>Advanced models, evaluations and analysis of the results (milestone II) (30%)</u>
  - What options/alternatives did you consider? (models, parameters, optimizers, regularizers, etc.)
  - Which options did you try and why?
  - Did you try anything creative? Please explain your idea.
  - How did you evaluate these models?
  - What did you conclude?

I considered Gaussian Discriminant analysis and SVM but the took to long to run so I used linear discriminant analyses and decision trees.


Receiver Operating Characteristic - LDA
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Fold 1 LDA (AUC = 0.85)
Fold 2 LDA (AUC = 0.88)
Fold 3 LDA (AUC = 0.87)
Fold 4 LDA (AUC = 0.77)
Fold 5 LDA (AUC = 0.51)
Mean LDA (AUC = 0.78 ± 0.14)
Chance


Receiver Operating Characteristic - Decision Tree
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Fold 1 Decision Tree (AUC = 0.55)
Fold 2 Decision Tree (AUC = 0.60)
Fold 3 Decision Tree (AUC = 0.62)
Fold 4 Decision Tree (AUC = 0.57)
Fold 5 Decision Tree (AUC = 0.43)
Mean Decision Tree (AUC = 0.55 ± 0.06)
Chance


Precision-Recall Curve - LDA
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Fold 1 LDA (AUC = 0.96)
Fold 2 LDA (AUC = 0.97)
Fold 3 LDA (AUC = 0.97)
Fold 4 LDA (AUC = 0.94)
Fold 5 LDA (AUC = 0.85)
Mean LDA (AUC = 0.94 ± 0.04)



I found out decision tree is particularly bad with

Decision Tree Mean AUC : 0.55

Decision Tree Mean F1: 0.8094133830039187

and time
Decision Tree Training Time: 15.742506742477417 seconds

- Analysis of feature engineering methods (15%)
  - What options did you consider?
  - Which option(s) did you try and why?
  - What was the effect of your method(s)?
- Analysis of model engineering methods (15%)
  - Analysis and results of hyperparameter tuning.
- Conclusion and summary of analysis (20%)

    - Comparison of models on test set.
    - Table(s) including all performance and resource utilization metrics and results to pick the best model.
    - Conclusion and summary; choice of a method and explanation of considerations in making the decision.
- Any references and links you have used to create or guide your work.

Group work

Include the following in **group.docx**:

1- Complete this declaration by adding your names:

We, ------------- (mention your names), declare that the attached assignment is our own work in accordance with the Seneca Academic Policy. We have not copied any part of this assignment, manually or electronically, from any other source including web sites, unless specified as references. We have not distributed our work to other students.

2- Specify what each member has done towards the completion of this work:

	Name	Task(s)- please **be specific. What would you include in your resume?**
1		
2		
3		

Update Your Resume

Showcase your experience for the employers!! Include your assignment in your resume. Describe the problem and your approach to solving it. Include a link to your GitHub or project webpage, if applicable.