# Machine Learning Project Plan

This document outlines the steps for the Machine Learning project plan. Each step is listed in a table with additional columns for Points, Self Grade, and Grader Grade. On UB Learns, upload a completed version of this sheet. The project is due 12/18/2024 @11:59 PM. No late submissions. I will ignore all emails about late submissions. **You MUST keep your apps running for one week, so you can turn it off 12/25/2024 @11:59PM.**

URL to your Jupyter Book:

<https://profit188.github.io/Income_notebook/content/index.html>

|  |  |  |  |
| --- | --- | --- | --- |
| **Step** | **Points** | **Self Grade** | **Grader Grade** |
| Create a normalized database (3NF). | 20 | 20 |  |
| Write SQL join statement to fetch data from the database and into Pandas DataFrame. | 5 | 5 |  |
| Explore the data to determine if you need to stratify it by some attribute when doing train/test split. Perform the train/test split. | 10 | 10 |  |
| Explore the data using yprofile and correlation matrix. Make observations about features, distributions, capped values, and missing values. Create a list of data cleanup tasks. | 10 | 10 |  |
| Experiment #1: Create a pipeline for preprocessing (StandardScaler, MinMaxScaler, LogTransformation, OneHotEncoding) and Logistic Regression. Log F1-score/(TP,TN,FN,FP) in MLFlow on DagsHub. – Cross validation 3/10 folds. Results—mean/std of CV results and results on the whole training data – add in parameter hyper tuning | 20 | 20 |  |
| Experiment #2: Create a pipeline for preprocessing and use LogisticRegression, RidgeClassifier, RandomForestClassifier, and XGBClassifier. Log results in MLFlow on DagsHub. | 10 | 10 |  |
| Experiment #3: Perform feature engineering and attribute combination. Log results in MLFlow. | 10 | 10 |  |
| Experiment #4: Perform feature selection using Correlation Threshold, Feature Importance, and Variance Threshold. Log results in MLFlow. | 10 | 10 |  |
| Experiment #5: Use PCA for dimensionality reduction on all the features. Create a scree plot to show which components will be selected for classification. Log results in MLFlow. | 10 | 10 |  |
| Experiment #6: Design and execute a custom experiment. Log results in MLFlow. | 10 | 10 |  |
| Experiment #7: Design and execute another custom experiment. Log results in MLFlow. | 10 | 10 |  |
| Create meaningful F1-score plots to compare experiments and determine the best model. | 10 | 10 |  |
| Save the final model using joblib.  Create a FastAPI application to serve the model.  Containerize the FastAPI application using Docker and push to Docker Hub.  Deploy the containerized API to a cloud platform. | 20 | 20 |  |
| Create a Streamlit app to interact with the deployed model for real-time classification. | 15 | 15 |  |
| Create a 12–15-minute video explaining the project. Walk through the Jupyter Notebook explaining your project. Your presentation should be coherent, organized, and you must speak clearly. **-5 points** for not having video with face. Use Zoom to record your video. Maybe you can record to cloud to link your video or you can upload to UB Box. | 30 | 30 |  |
| Create a JupyterBook website with your resume, embedded video, and final code. Include links to MLFlow/DagsHub experiments, Docker Hub container, deployed model, and Streamlit app. | 20 | 20 |  |
| **Grade** | 220 | 220 |  |

Use the following section to inform your grader why you missed certain things or how they were not applicable to your project so your grades can be adjusted.

* All links are added in notebook useful links