Machine Learning| Project Overview

Project

- Groups of 4-5
- Same data & target as the sklearn / credit Jupyter notebook example
- Task 1: build as good of a model as you can using one or several non-linear techniques of your choice: Boosted Trees (i.e. XGBoost), random forest, or MLP
    - Evaluate models using cross-validation & metrics of your choice (with justification); compare these to the performance of the baseline model from the tutorial
- Task 2: Model evaluation & strategy creation
    - Evaluate your model based on well-reasoned criteria as discussed in class, and discuss its performance
    - Based on well-reasoned assumptions of benefit of onboarding new customers, and costs of onboarding "bad" customers, devise a strategy by constructing a profit curve and choosing the optimal cut-off for your strategy

Group & Individual Component

Group component (groups of 5) – **30 points total**:

1. Document your data transformations (**15 pts**):
    - Data dictionary for the final feature tables (5 pts)
    - Documentation on how you got there (include code, but should be readable without looking at the code) (5 pts)
    - Documentation of QA (5 pts)
2. Document models (**15 pts**)
    - Code to produce (5 pts)
    - Appropriate performance measures + rationale on why these were chosen (5 pts)
    - Searching for the best model: approaches, hyperparameter tuning, evaluation loss, etc. (5 pts)

This section should demonstrate a technical understanding of course concepts. Focus on being technically correct vs "fancy", but feel free to explore as well. This can be an expansion of the notebook from class, but feel free to see other tools outside of Jupyter you find helpful.

Group & Individual Component

Individual component – **30 points total**

1. Individual business write up (**10 pts**):
    - Clearly explain business value of findings and how findings can be applied, including strategy changes and deployment of models; ideally this will involve constructing a profit curve to determine your optimal strategy, but other explanations for value of your model are also valid (5 pts)
    - Common-sense / intuitive explanations for feature engineering, best features, other insights (5 pts)
2. Individual technical write up (**10 pts**):
    - Explain model validation & approaches chosen, how you chose the best model and its hyperparameters, why you will trust it in production (5 pts)
    - Explain what you would do given more time to improve on your best model (5 pts)
3. Outline next steps – technical & business (**10 pts**): for this question, include a prompt that can use to draft an initial response, the response generated by the prompt, summarize your prompt engineering response and include the initial response generated + your final edited response.
    - Imagine you are the new VP for Machine Learning in credit risk. What is your 2 year plan – what kind of data will you collect, what kind of techniques will you apply, what products can you create and how will that integrate into the overall business strategy, etc. (5 pts)
    - How will you monitor this model once it is in production – include considerations on changes in your input features, statistical model performance & business KPI's (5 pts)

Project Tips

- Don’t boil the ocean – you can work on these problems infinitely long
- 50-100 features for your model first model is fine, and then depending on your approach the number features should grow somewhat for the linear model (e.g. if you make multiple bins per feature)
- Think about bias / leakage and avoid it – your write up should clearly explains the business drivers that go into your models together with model validation; are any of your best features suspicious to you for any reason? Do you expect the model to generalize perfectly?
- Get creative, and feel free to include interesting you tried in the technical sections
- Showcase understanding of the course concepts and be clear