TDS Project: part 2- Advanced Model Analysis, Optimization and Conclusions

---

Objective

In the second part of the project, you will continue working on your data science pipeline. After choosing a dataset and applying an initial model, you will focus on enhancing model performance, analyzing the improved model, and drawing meaningful conclusions from the data exploration and results.

Steps to Follow

1. Error Analysis Conclusions & work plan

- Refer to the error analysis you previously conducted, and draw conclusions. I.e. what might have caused the errors, and what you plan to do in order to solve these issues.

2. Improving Model Performance

- Identify and address weaknesses in the initial baseline model.
- Perform techniques such as:
  - Hyperparameter tuning.
  - Feature engineering (creating new features, transforming existing ones).

        - Handling missing data or outliers effectively.
        - Balancing data (if applicable) for fairness or model accuracy.

IMPORTANT! Changing a learning algorithm doesn't count as an improvement!

3. Analyzing the Improved Model

- **Feature Importance**: Use tools such as SHAP, permutation importance, or model-specific importance metrics (e.g., for Random Forest).
- **Explain Model Performance**: Compare results of the improved model against the baseline and explain why it performs better.
    - Metrics to use: R-squared, pearson correlation, ROC-AUC, RMSE, etc.
- **Visualization**: Create clear plots to highlight model performance improvements and the role of features.

---

4. Drawing Conclusions About the Data & Creative Applications

- Analyze what the improved model reveals about the data:
    - **Feature Significance**: Which features are most influential?
    - **Biases or Trends**: Are there identifiable patterns or biases (e.g., feature correlation with outcomes)?

    - **Data Insights**: Any surprising findings? Are specific features driving predictions more than expected?
- Explain findings in a user-friendly narrative to communicate insights effectively.
- Propose potential applications for the dataset or findings. Examples include:
    - Real-world use cases based on trends (e.g., targeting marketing campaigns, improving credit scoring systems).
    - Building predictive tools for relevant stakeholders.
    - Ideas inspired by external sources: notebooks, papers, or applications (with proper citations).

Deliverables

Assignment due: 21.12.24

You will create a new notebook, where you will begin by loading the model from part 1 followed by the analysis. Afterwards, the new pipeline and the rest should be presented in snippets and markdowns.

- **Notebook Presentation**:
    - Include clear **markdown descriptions**, **visualizations**, and short **code blocks** that showcase the following:
        - Your full workflow: model improvement, evaluation, and conclusions.
        - Side-by-side comparison of baseline and improved models.

        - Visual explanations of feature importance and model performance.
- **Conclusions**:
    - Summarize key findings and highlight actionable insights.
    - Justify the improvements made and their impact on performance.

By following this structured approach, you will showcase a refined, results-driven workflow while drawing meaningful insights from the data in an organized and user-friendly manner.

For special requests/approvals, email me- itay.elyashiv@gmail.com