Prediction of Loan Default

Northwestern Mutual Interview Exercise for Data Scientist Applicants

Purpose: We want to evaluate a data science applicant on their ability to:

- Form original and thoughtful ideas guided by the data provided.
- Produce working, clean code.
- Demonstrate capabilities of using data wrangling, statistical, and machine learning tools in Python.
- Creatively explore multiple modeling approaches.
- Recommend, using appropriate metrics, why their model would appropriately meet the goals of the problem at hand.

Requirements: We expect a Jupyter Notebook written in Python complete with the entire thought process of the applicant from project understanding and data exploration to a final predictive model. All code submitted by the applicant should be the candidate's own original work, with the possible exception of small code snippets from Stack Overflow or other references.

Problem: A bank is looking to accelerate their underwriting process for home equity lines of credit. One way of doing this is by using predictive models trained on recent loan underwriting data to predict individuals of having low or high risk of defaulting on their loan. You are tasked with creating such a model and conveying to the stakeholder (the bank) the pros and cons of various modeling techniques in the context of model performance and interpretability/explainability.

Required Sections:

1. Business Understanding – Clearly state the objective, as well as other considerations or limitations of the data when considering the problem at hand.
2. Data Exploration – Display thoughtful use of data visualization tools and summary statistics.
3. Data Wrangling – Use appropriate data wrangling tools to correctly mold the data for the multiple models that will be tested.

4. Modeling – Illustrate the effectiveness of different modeling algorithms and/or approaches.
5. Assessment – Use appropriate metrics for assessing and comparing models with each other. Explain why one model is superior considering performance metrics and a business use case.
6. Summary – Provide a short summary of the findings and proposed solution. This can be in either a document or presentation.

Data: *NMLoanDefault.csv*

Data Dictionary:

Variable	Description
PROPERTY_VALUE_AMT	Value of current property
TARGET	0 = loan repaid, 1 = loan defaulted
CRDT_LINE_CNT	Number of credit lines
DEROG_CNT	Number of derogatory reports
DEBT_INC_RTIO_AMT	Debt to income ratio
LOAN_AMT	Amount of loan requested
REASON_CDE	Reason for the loan: DebtCon = Debt Consolidation, HomeImp = Home Improvement
YOJ_AMT	Years at present job
MORTGAGE_DUE_AMT	Amount due on existing mortgage
RCNT_CRDT_CNT	Number of recent credit lines
OLD_AGE_TRADE_AMT	Age of oldest trade line in months
JOB_CDE	Occupation category
DELINGQ_CNT	Number of delinquent credit lines