Data Science and Machine Learning Assignment WiSe 25/26

FinTech Dataset

Dataset Overview

This dataset contains customer information from a European FinTech startup that offers various financial products including banking, credit cards, and investment services. The dataset is designed for educational purposes to practice multiple machine learning tasks within a realistic business context.

Dataset Specifications:

- **Size:** 8,662 customers × 23 features
- **Format:** CSV with header row

Business Problem

The FinTech startup faces several analytical challenges that require different machine learning approaches:

1. **Customer Churn Prediction:** Identify customers who are likely to leave the platform
2. **Customer Lifetime Value Prediction:** Estimate the monetary value each customer will generate
3. **Customer Segmentation:** Group customers into distinct segments for targeted marketing strategies

Dataset Features

Identifier

- **Customer_ID:** Unique customer identifier

Customer Demographics

- **CCreditScore:** Customer's credit score (numeric)
- **CGeography:** Customer's country of residence (France, Germany, Spain)

- **CGender:** Customer's biological gender (Male, Female)
- **CAge:** Customer's age in years (numeric)
- **CTenure:** Number of years as a registered customer (numeric)
- **CBalance:** Current account balance (numeric, can be negative due to authorized overdrafts)
- **CNumOfProducts:** Number of financial products owned (1-4)
- **CHasCrCard:** Whether the customer has a credit card (0/1)
- **CIsActiveMember:** Whether the customer is an active member (0/1)
- **CEstimatedSalary:** Estimated annual salary (numeric)

Temporal & Behavioral Features

- **Days_Since_Onboarding:** Days since customer registration (redundant with Account_Age_Months; recommend using Account_Age_Months instead)
- **Account_Age_Months:** Account age in months (preferred temporal feature - perfect correlation with Days_Since_Onboarding)
- **Avg_Monthly_Transactions:** Average number of transactions per month over customer lifetime
- **Transaction_Variance:** Variance in monthly transaction counts (measures consistency of customer activity)
- **Last_Login_Days_Ago:** Days since last platform login (note: weak correlation with churn - many active users churn and inactive users remain)
- **Support_Tickets_Count:** Total number of customer support interactions over customer lifetime
- **Mobile_App_Usage_Hours:** Average monthly mobile app usage in hours
- **Onboarding_Month:** Month when customer joined (1-12, seasonal patterns may exist)
- **Is_Holiday_Onboarding:** Whether customer joined during holiday season (0/1)

Target Variables

1. Churn (Binary Classification)

- **Description:** Whether the customer has left the platform
- **Values:** 0 = No Churn, 1 = Churn
- **Distribution:** Imbalanced (majority No Churn)

2. CLV_Continuous (Regression)

- **Description:** Customer Lifetime Value in monetary terms
- **Values:** Continuous numeric values in dollars
- **Range:** Varies significantly across customers

Classification Tasks

Task 1: Data Preprocessing

Before you can build your models, prepare the dataset for analysis.

- Apply your data preprocessing knowledge to thoroughly clean and prepare the dataset before training.

Task 2: Build and Evaluate Basic Models

Experiment with different machine learning models to predict customer churn (binary classification using the 'Churn' target variable).

- Use three different basic machine learning models from lecture 02 (e.g., Decision Trees, k-Nearest Neighbors, Naive Bayes, or Logistic Regression).
- Describe your approach to applying each model, including the assumptions and motivations behind each choice.
- Evaluate the performance of the models.
- Justify which machine learning model you would choose based on the performance results from this task.

Note: No hyperparameter tuning is expected at this stage.

Task 3: Optimize the Selected Model

Optimize the best-performing model from Task 2 through hyperparameter tuning.

- Optimize the performance using advanced techniques from lecture 04 (Note: Choose the model that performed best in Task 2). Be sure to analyze whether your model is overfitted or underfitted and use cross-validation for evaluation.
- Describe your approach to tuning the model. Especially, specify the chosen hyperparameters and why that.
- Interpret the performance of the model after optimization and compare it to the initial model from Task 2.
- Justify which final machine learning model you would choose after optimization.

Task 4: Apply an Ensemble Learning Technique

Ensemble learning combines multiple models to improve overall performance. In this task, you will explore whether using an ensemble technique improves your performance.

- Use one ensemble learning technique (from lecture 05). Explain and describe your approach. Evaluate whether this approach leads to better performance than the models from Tasks 2 and 3. Be sure to analyze if the ensemble model is overfitted or underfitted.
- Describe your approach to using and tuning the ensemble model, including the hyperparameters.
- Compare the performance of the ensemble model to the best model from Task 3 and use appropriate evaluation metrics.

General Guidelines

- **Notebook Structure:** Ensure your notebook is well-organized, with clear sections for each task. Use markdown cells to provide explanations and commentary throughout your work.
- **Code Documentation:** Your code should be clean, well-documented, and easy to follow. Make use of comments to explain key steps and decision points in your analysis.

Closing Thoughts

This assignment gives you the opportunity to explore a wide range of classification techniques and understand their real-world applications. Your final goal is to understand how each technique performs and how you can use data science to provide meaningful insights for classification problems.

Regression Tasks

Task 1: Data Preprocessing

Start to clean and preprocess the dataset to ensure it is ready for analysis.

- Use your knowledge of data preprocessing to thoroughly clean and prepare the dataset for training.
- **Note:** You can reuse or build upon the data cleaning steps from the classification assignment. A single, well-documented preprocessing script or notebook is encouraged.

Task 2: Train and Evaluate a Multiple Regression Model

- Build a multiple linear regression model to predict Customer Lifetime Value (regression using the 'CLV_Continuous' target variable)".
- Describe your approach to building the model and interpret the performance based on appropriate evaluation metrics.

Note: You are not expected to use regularization or hyperparameter tuning at this stage.

Task 3: Polynomial Regression Models

- Train and evaluate at least two polynomial regression models with different polynomial degrees.
- Compare the results of the polynomial regression models with the linear regression model from Task 2. Discuss how the model's performance changes as the degree of the polynomial increases.

Task 4: Model Comparison

- Compare the performance of the multiple linear regression model from Task 2 and the polynomial regression models from Task 3.
- Discuss which model performs better and provide a rationale for your conclusions.

Note: Regularization and hyperparameter tuning are not required at this point.

Task 5: Regularization and Optimization

- Optimize your results by applying regularization techniques, such as Ridge or Lasso regression (Note: select the best-performing model in Task 4).

- Use cross-validation and hyperparameter tuning (e.g., GridSearchCV or RandomizedSearchCV) to further optimize your model's performance.
- Explain whether your model is overfitting or underfitting, and interpret the results using evaluation metrics and visualizations.

Task 6: Regression Tree Model

- Train and evaluate a regression tree model to predict Customer Lifetime Value (regression using the 'CLV_Continuous' target variable).
- Use hyperparameter tuning and cross-validation to optimize the regression tree model.
- Analyze and explain whether your model is overfitting or underfitting, and interpret the results with appropriate metrics and plots.

Task 7: Model Comparison

- Compare the performance of the machine learning model optimized in Task 5 and the regression tree model from Task 6.
- Discuss which model performs better. Also explain why the model performs better. Use metrics and reasoning based on the strengths and weaknesses of each model.

Task 8: Ensemble Learning

- Use an ensemble learning technique to improve your model's performance. Compare the results with the model you selected from Task 7.
- Use hyperparameter tuning and cross-validation to optimize your ensemble model.
- Analyze whether your ensemble model is overfitting or underfitting, and explain the results based on your evaluation.

Guidelines and Tips

- **Notebook Structure:** Ensure your notebook is well-organized, with clear sections for each task. Use markdown cells to explain your process, observations, and conclusions throughout.
- **Code Documentation:** Your code should be clean and well-commented to make it easy to follow. Clear documentation will help to understand your approach.

Closing Thoughts

This assignment gives you the opportunity to explore a wide range of regression techniques and understand their real-world applications. Your final goal is to understand how each technique performs and how you can use data science to provide meaningful insights.

Clustering Tasks

Task 1: Data Preprocessing

Clean and prepare your data.

- Use appropriate preprocessing techniques to ensure the dataset is ready for clustering.
- **Note:** You can reuse or build upon the data cleaning steps from the previous assignments.

Task 2: Clustering Analysis

Use the preprocessed data and clustering algorithms to identify distinct customer groups.

- Use clustering to segment your customers into meaningful groups.
- Determine the optimal number of clusters and evaluate the quality of your clusters using appropriate metrics.

Task 3: Cluster Profiling

Focus on the characteristics of each group.

- Analyze and describe each cluster by examining the characteristics of the customers within the clusters.
- Compare the clusters and highlight key differences and similarities.
- Visualize your clusters to present the differences.

Task 4: Marketing Recommendations

Translate these findings into real-world business strategies.

- Propose actionable marketing strategies for each customer segment. Think about how the marketing team can create targeted campaigns that resonate with each group.

Guidelines

- **Notebook Structure:** Ensure your notebook is well-organized with clear sections for each task. Use markdown cells to provide explanations, instructions, and comments throughout.
- **Code Documentation:** Your code should be clean and well-documented. Add comments to explain key steps, so that anyone reviewing your notebook can easily follow your thought process.

- **Visualization:** Use visualizations wherever applicable to help tell the story of your data. Plots can be a powerful tool to illustrate your findings and recommendations.

Final Thoughts

This assignment is about thinking critically and creatively. Take this opportunity to explore the data, experiment with clustering techniques, and provide insights that could have a real impact on business decisions.

Good luck, and have fun with your analysis!