Project Journal

Project Title: *Data-Driven Insights for Banking: Financial and Consumer Behavior Analysis*

Team Members Name:

1. Gonzalo Mauricio Bugueño Cuadra
2. Bhagyam Babu
3. Sneha Mini Biju
4. Suryakarthik

Development

Member: Bhagyam Babu

Week 1: Preliminary Research and Data Display

This week's emphasis was on comprehending the dataset and using visualizations to examine its properties.

Histogram: Used Matplotlib to plot a histogram for the LIMIT_BAL column.

Observation: Most clients (approx. 3,400) had a loan limit between 0 and 1 lakh TWD.

Boxplot: Created a boxplot using Seaborn to analyse LIMIT_BAL by payment defaults.

Key Insight: Non-defaulters generally had higher credit limits compared to defaulters.

Scatterplot: Analysed the relationship between age (AGE) and credit amount (LIMIT_BAL) while factoring in default status. Finding: Lower credit limits correlated with higher default rates, irrespective of age.

Time Spent: 7hours

During graphing, overlapping variable names caused confusion.

Renaming variables for clarity and documenting naming conventions were the solutions or actions implemented.

Week 2: Model Training and Evaluation.

Random Forest Classifier:

To determine if a client will default, a Random Forest Classifier was constructed.

Set up to balance overfitting and complexity with a maximum depth of 30 and 100 estimators.

Obtained an accuracy score of 81.74%, which indicates that 82 out of 100 predictions were correctly identified by the model.

Observations:

The model did a respectable job of catching important trends in the data.

emphasized how crucial feature selection and tuning are to improved performance.

Time Spent: 9 hours

On the test set, the model's predictions were erroneous due to overfitting on the training data.

Improved generalization on unknown data by lowering the maximum depth of the trees and modifying other hyperparameters to avoid overfitting.

Week 3: Database Integration

For organized storage and additional analysis, the dataset was integrated into a SQLite database.

Database Setup:

Created a project.db file and defined a table (defaultd) with schema corresponding to the dataset features.

Data Insertion:

Added rows to the database table from the dataset (X and Y),well-defined schema restrictions that guaranteed data integrity and consistency.

Outcome: Successfully migrated data into a relational format for streamlined querying and analysis.

Time Spent: 5hours

Errors were induced by data insertion.Verified rows and standardized data types prior to inclusion.

Week 4: Model Serialization and Reusability

Focused on ensuring the trained model can be reused efficiently.

Model Serialization:

The trained Random Forest model (rf) was saved as classifier.pkl using the pickle module.

This saves time and computational resources by enabling future use without retraining.

Directory Management:

Organized files by saving the serialized model in a dedicated pickled directory.

Reflection:

In practical applications, serialization makes machine learning models more useful.

Time Spent: 2hours

Total Time Spent: 23 hours

Both data analysis and model development saw notable advancements throughout the project.

To investigate the connections between credit limit, age, and payment default behaviour, visualizations including histograms, boxplots, and scatterplots were made in Week 1. The results showed that lower credit limits were associated with greater default rates, whereas non-defaulters generally had higher credit limits. A Random Forest Classifier with an accuracy of 81.74% was created in Week 2 to forecast client defaults. To improve generalization on test data, efforts were directed toward lowering overfitting through hyperparameter adjustments. The dataset was successfully integrated into a SQLite database in week three, enabling organized storage and effective querying. Finally, in Week 4, the pickle module was used to serialize the trained Random Forest model for reuse, saving computing power for subsequent forecasts. Additionally, files were arranged to make the serialized model easily accessible. Additionally, the team was able to debate problems and exchange ideas through regular communication and teamwork, which made sure that solutions were applied as a group. The team was able to proceed effectively and sustain a high level of work throughout the project by utilizing one another's abilities and cooperating to overcome challenges.

Project related meetings :4 hours

Dataset Research: 2hours

Report: 5hours (Reading related research papers)

Presentation