10 Academy: Artificial Intelligence Mastery

Week 6 Challenge Document

Date: 22 Jan - 28 Jan 2025

Overview

Business Need

You are an Analytics Engineer at **Bati Bank**, a leading financial service provider with over 10 years of experience. Bati Bank is partnering with an upcoming successful eCommerce company to enable a **buy-now-pay-later** service - to provide customers with the ability to buy products by credit if they qualify for the service. You are assigned a project to create a **Credit Scoring Model** using the data provided by the eCommerce platform.

Credit scoring is the term used to describe the process of assigning a quantitative measure to a potential borrower as an estimate of how likely the default will happen in the future. Traditionally, creditors build credit scoring models using statistical techniques to analyze various information of previous borrowers in relation to their loan performance. Afterward, the model can be used to evaluate a potential borrower who applies for a loan by providing the similar information which has been used to build the model. The result is either a score which represents the creditworthiness of an applicant or a prediction of whether an applicant will default in the future.

The definition of default in the context of credit scoring may vary between each financial institution as long as it complies with the Basel II Capital Accord - you must read this reference to understand the factors the bank needs to take into account to start a new loan procedure. A quick summary of the Basel II Capital Accord can be found in this reference.

Your job is to build a product that does the following

1. Defines a proxy variable that can be used to categorize users as high risk (bad) or low risk (good)
2. Select observable features that are good predictors (have high correlation) of the default variable defined in 1)
3. Develop a model that assigns risk probability for a new customer
4. Develop a model that assigns credit score from risk probability estimates
5. Develop a model that predicts the optimal amount and duration of the loan

Data and Features

The data for this challenge can be found here. Or you can find it also here: Xente Challenge | Kaggle

Data fields

- **TransactionId:** Unique transaction identifier on platform
- **BatchId:** Unique number assigned to a batch of transactions for processing
- **AccountId:** Unique number identifying the customer on platform
- **SubscriptionId:** Unique number identifying the customer subscription
- **CustomerId:** Unique identifier attached to Account
- **CurrencyCode:** Country currency
- **CountryCode:** Numerical geographical code of country
- **ProviderId:** Source provider of Item bought.
- **ProductId:** Item name being bought.
- **ProductCategory:** ProductIds are organized into these broader product categories.
- **ChannelId:** Identifies if customer used web,Android, IOS, pay later or checkout.
- **Amount:** Value of the transaction. Positive for debits from customer account and negative for credit into cus...
- **Value:** Absolute value of the amount
- **TransactionStartTime:** Transaction start time
- **PricingStrategy:** Category of Xente's pricing structure for merchants
- **FraudResult:** Fraud status of transaction 1 -yes or 0-No

Learning Outcomes

Skills:

- Advanced use of scikit-learn
- Feature Engineering
- ML Model building and fine-tuning
- CI/CD deployment of ML models
- Python logging
- Unit testing
- Model management
- MLOps with CML, and MLFlow

Knowledge:

- Reasoning with business context
- Data exploration
- Predictive analysis
- Machine learning
- Hyperparameter tuning

- Model comparison & selection

Communication:

- Reporting on statistically complex issues

Competency Mapping

The tasks you will carry out in this week's challenge will contribute differently to the 11 competencies 10 Academy identified as essential for job preparedness in the field of Data Engineering, and Machine Learning engineering. The mapping below shows the change (lift) one can obtain through delivering the highest performance in these tasks.

Competency	**Potential contributions from this week**
Professionalism for a global-level job	Articulating business values
Collaboration and Communicating	Reporting to stakeholders
Software Development Frameworks	Using Github for CI/CD, writing modular codes, and packaging
Python Programming	Advanced use of python modules such as Pandas, Matplotlib, Numpy, Scikit-learn, Prophet and other relevant python packages
SQL programming	MySQL db create, read, and write
Data & Analytics Engineering	data filtering, data transformation, and data warehouse management
MLOps & AutoML	Pipeline design, data and model versioning,
Deep Learning and Machine Learning	NLP, topic modelling, sentiment analysis

Team

Tutors:

- Mahlet
- Elias
- Rediet
- Kerod
- Emtinan
- Rehmet

Key Dates

- Discussion on the case - 09:00 UTC on Wednesday 22 Jan 2025. Use #all-week6 to pre-ask questions.
- Interim Solution - 20:00 UTC on Friday 24 Jan 2025.
- Final Submission - 20:00 UTC on Tuesday 28 Jan 2025

Deliverables

Task 1 - Understanding Credit Risk

Focus on understanding the concept of Credit Risk

Key references

- https://www3.stat.sinica.edu.tw/statistica/oldpdf/A28n535.pdf
- https://www.hkma.gov.hk/media/eng/doc/key-functions/financial-infrastructure/alternative_credit_scoring.pdf
- https://thedocs.worldbank.org/en/doc/935891585869698451-0130022020/original/CREDITSCORINGAPPROACHESGUIDELINESFINALWEB.pdf
- https://towardsdatascience.com/how-to-develop-a-credit-risk-model-and-scorecard-91335fc01f03
- https://corporatefinanceinstitute.com/resources/commercial-lending/credit-risk/
- https://www.risk-officer.com/Credit_Risk.htm

Task 2 - Exploratory Data Analysis (EDA)

1. **Overview of the Data:**
   - Understand the structure of the dataset, including the number of rows, columns, and data types.
2. **Summary Statistics**
   - Understand the central tendency, dispersion, and shape of the dataset's distribution.
3. **Distribution of Numerical Features**
   - Visualize the distribution of numerical features to identify patterns, skewness, and potential outliers.
4. **Distribution of Categorical Features**
   - Analyzing the distribution of categorical features provides insights into the frequency and variability of categories.
5. **Correlation Analysis**
   - Understanding the relationship between numerical features.
6. **Identifying Missing Values**
   - Identify missing values to determine missing data and decide on appropriate imputation strategies.
7. **Outlier Detection**
   - Use box plots to identify outliers.

Task 3 - Feature Engineering

1. **Create Aggregate Features**
   Example:
   - **Total Transaction Amount:** Sum of all transaction amounts for each customer.
   - **Average Transaction Amount**: Average transaction amount per customer.
   - **Transaction Count:** Number of transactions per customer.
   - **Standard Deviation of Transaction Amounts:** Variability of transaction amounts per customer.
2. **Extract Features**
   Example:
   - **Transaction Hour:** The hour of the day when the transaction occurred.
   - **Transaction Day:** The day of the month when the transaction occurred.
   - **Transaction Month:** The month when the transaction occurred.
   - **Transaction Year:** The year when the transaction occurred.
3. **Encode Categorical Variables**
   Convert categorical variables into numerical format by using:
   - **One-Hot Encoding:** Converts categorical values into binary vectors.
   - **Label Encoding:** Assigns a unique integer to each category.
4. **Handle Missing Values**
   Use imputation or Removal to handle missing values
   - **Imputation**: Filling missing values with mean, median, mode, or using more methods like KNN imputation.
   - **Removal**: Removing rows or columns with missing values if they are few.
5. **Normalize/Standardize Numerical Features**
   Normalization and standardization are scaling techniques used to bring all numerical features onto a similar scale.
   - **Normalization**: Scales the data to a range of [0, 1].
   - **Standardization**: Scales the data to have a mean of 0 and a standard deviation of 1.

Feature Engineering using:

- https://pypi.org/project/xverse/
- https://pypi.org/project/woe/
- WEIGHT OF EVIDENCE (WOE) AND INFORMATION VALUE (IV) EXPLAINED

Task 4 - Default estimator and WoE binning

The purpose of any credit scoring system is to classify users as high risk or low-risk. High-risk groups are those with high likelihood of default - those who do not pay the loan principal and interest in the specified time frame.

To simplify the process, here we want to construct a variable based on RFMS formalism that classifies users into good (high RFMS score) and bad (low RFMS score). You may use this reference to help you connect the dots.

1. Construct a default estimator (proxy)
    a. By visualizing all transactions in the RFMS space, establish a boundary where users are classified as high and low RFMS scores.
    b. Assign all users the good and bad label
2. Perform Weight of Evidence (WoE) binning following this, this or this references

Task 5 - Modelling

1. **Model Selection and Training**
    a. **Split the Data**
    Splitting the data into training and testing sets helps evaluate the model's performance on unseen data.
    b. **Choose Models**
    Choose at least two models from the following:
        - Logistic Regression
        - Decision Trees
            - Random Forest
        - Gradient Boosting Machines (GBM)
    c. **Train the Models**
    Train the models on the training data
    d. **Hyperparameter Tunning**
    Improve model performance using hyperparameter tuning, use techniques like:
        - Grid Search
        - Random Search
2. **Model Evaluation**
Assess model performance using the following metrics
    a. **Accuracy**: The ratio of correctly predicted observations to the total observations.
    b. **Precision:** The ratio of correctly predicted positive observations to the total predicted positives.
    c. **Recall (Sensitivity):** The ratio of correctly predicted positive observations to all observations in the actual class.
    d. **F1 Score:** The weighted average of Precision and Recall.
    e. **ROC-AUC:** Area Under the Receiver Operating Characteristic Curve, which measures the ability of the model to distinguish between classes.

Task 6 - Model Serving API Call

- Create a REST API to serve the trained machine-learning models for real-time predictions.
  - **Choose a framework:**
    - Select a suitable framework for building REST APIs (e.g., Flask, FastAPI, Django REST framework).
  - **Load the model:**
    - Use the model from Task 4 to load the trained machine-learning model.
  - **Define API endpoints:**
    - Create API endpoints that accept input data and return predictions.
  - **Handle requests:**
    - Implement logic to receive input data, preprocess it, and make predictions using the loaded model.
  - **Return predictions:**
    - Format the predictions and return them as a response to the API call.
  - **Deployment:**
    - Deploy the API to a web server or cloud platform.

Tutorials Schedule

Overview

In the following, the colour **purple** indicates morning sessions, and **blue** indicates afternoon sessions.

Wednesday:

- Introduction to the challenge(Mahlet)
- Introduction to Credit Risk Analysis and Modeling (Kerod)

Thursday:

- Feature Engineering, WEIGHT OF EVIDENCE(WoE), and INFORMATION VALUE (IV) (Elias)

Friday:

- Model Training, Hyperparameter Tuning, and Evaluation (Emtinan)
- Model Serving and Deployment (Rediet)

Interim Submission

- A review report of your reading and understanding of Task 1 and any progress you made in other tasks.
- Link to your GitHub.

Feedback

You may not receive detailed comments on your interim submission but will receive a grade.

Final Submission

- A blog post entry (which you can submit for example to Medium publishing) or a pdf report.
- Link to your Github code, and make sure to include screenshots demonstrating anything else you have done.

Feedback

You will receive comments/feedback in addition to a grade.

References

1. Loss functions
2. Sklearn pipelines
3. Merging dataframes
4. RandomForests

Credit Risk

1. https://www.investopedia.com/terms/c/creditrisk.asp
2. https://investopedia.com/terms/c/creditspread.asp
3. https://cleartax.in/glossary/credit-risk/
4. https://corporatefinanceinstitute.com/resources/commercial-lending/credit-risk/
5. https://www.risk-officer.com/Credit_Risk.htm

   Publications:

   a. Credit Risk Determinants in Selected Ethiopian Commercial Banks: A Panel Data Analysis

   b. Factors Affecting Credit Risk Exposure of Commercial Banks in Ethiopia: An Empirical Analysis

   c. Credit Risk Analysis of Ethiopian Banks: A Fixed Effect Panel Data Model.
6. https://drive.google.com/drive/folders/1pAXmJ_SI46D4Ex-nV0pDGvpxa7HD5erW?usp=drive_link
7. https://shichen.name/scorecard/

MLOps

1. Auto-sklearn — AutoSklearn 0.12.7 documentation (automl.github.io)
2. https://machinelearningmastery.com/hyperparameter-optimization-with-random-search-and-grid-search/
3. Hyperparameter tuning. Grid search and random search
4. https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/

Kaggle kernels

1. https://www.kaggle.com/datasets/atwine/xente-challenge

Feature Engineering in Credit Scoring

1. https://pypi.org/project/xverse/
2. https://pypi.org/project/woe/
3. https://github.com/JGFuentesC/woe_credit_scoring
4. https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html
5. https://shichen.name/scorecard/

Related Optional References

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8860138/
- Credit Risk Determinants in Selected Ethiopian Commercial Banks: A Panel Data Analysis
- Factors Affecting Credit Risk Exposure of Commercial Banks in Ethiopia: An Empirical Analysis
- Credit Risk Analysis of Ethiopian Banks: A Fixed Effect Panel Data Model.