Presentation: The Importance of Such a Project in the Banking and Finance Sector

Introduction

In the banking and finance sector, data management and its exploitation through artificial intelligence (AI) models have become strategic priorities. This supervised classification project, based on multiple machine learning models, offers a versatile approach to solving critical problems. Thanks to its automation and ability to deploy the optimal model, it can deliver significant value in this highly competitive industry.

Applications in Banking and Finance

1. **Risk Analysis and Management**
   - **Objective**: Identify clients with a high risk of default or fraud.
   - **Example**: This project can classify loan applications based on risk levels (low, medium, high) using client data (income, banking history, payment behavior).
   - **Impact**:
     - Reduced financial losses.
     - Improved decision-making accuracy for loan approvals.
2. **Fraud Detection**
   - **Objective**: Detect fraudulent transactions in real-time.
   - **Example**: Classify transactions as legitimate or fraudulent using features such as transaction amount, location, or account history.
   - **Impact**:
     - Early detection of fraud reduces losses.
     - Protection of client assets and the institution’s reputation.
3. **Customer Segmentation**
   - **Objective**: Group customers into homogeneous segments to better target products and services.
   - **Example**: Classify customers based on their investment preferences, savings behavior, or propensity to adopt new products.
   - **Impact**:
     - Personalization of services and increased customer satisfaction.
     - Optimization of marketing campaigns.
4. **Portfolio Optimization**
   - **Objective**: Identify investment strategies tailored to different risk profiles.
   - **Example**: Classify financial products based on profitability and associated risks for specific clients.
   - **Impact**:
     - Better management of assets under management (AUM).
     - Reduced losses through precise risk assessment.
5. **Compliance and Regulation**
   - **Objective**: Identify activities that may not comply with current regulations.
   - **Example**: Classify transactions or bank accounts based on their compliance with anti-money laundering (AML) standards.
   - **Impact**:
     - Reduced fines and penalties for non-compliance.
     - Improved relationships with regulators.

Specific Advantages of the Project

1. **Model Comparison for Optimal Performance**
   - By automatically comparing classification models, the most suitable algorithm for a given problem (e.g., fraud detection or risk assessment) is selected, ensuring reliable and accurate predictions.
2. **Automation and Cost Reduction**

    - Automating the selection and deployment of the best model reduces costs associated with manual development, testing, and optimization.

3. **Flexibility and Adaptability**
    - The pipeline is designed to work with various types of data (transactional, demographic, financial), making it adaptable to multiple use cases in the financial sector.
4. **Accessibility via User Interface**
    - Streamlit makes the project accessible to non-technical financial analysts, allowing them to submit data and obtain predictions without writing code.
5. **Simple and Portable Deployment**
    - Containerization with Docker ensures the project can be quickly deployed on different infrastructures (local servers or cloud), making the tool accessible across all branches of a banking organization.

Global Impact on the Banking and Finance Sector

1. **Enhanced Decision-Making**
    - Accurate predictions on customer behavior or financial trends enable data-driven decisions, increasing profitability.
2. **Improved Customer Experience**
    - By better targeting customer needs, banks can personalize products, retain existing customers, and attract new investors.
3. **Reduction of Financial and Operational Risks**
    - By anticipating defaults or fraud, financial institutions can minimize potential losses and protect their reputation.
4. **Improved Regulatory Compliance**
    - By identifying and reporting suspicious or non-compliant transactions, institutions avoid costly sanctions.

Conclusion

This project is a strategic asset for the banking and finance sector, where precision and efficiency are paramount. By integrating best practices in machine learning, it addresses complex problems while reducing costs and improving decision-making.

This type of solution reflects a shift toward more data-driven and technology-focused finance, preparing institutions for the challenges and opportunities of tomorrow.

Specifications Document: Supervised Classification Application Based on Multiple Models

1. Project Context

In the field of data analysis, supervised classification is essential for predicting categories based on provided features. The primary goal is to develop an application capable of:

- Evaluating multiple classification models.
- Comparing their performance.
- Automating the selection and saving of the best-performing model for deployment.

This project will be made available to end-users through a user-friendly and containerized web interface.

2. Objectives

1. **Development of a classification pipeline:**
   - Implement and evaluate multiple supervised models (KNN, Random Forest, SVM, Logistic Regression, Naive Bayes).
   - Calculate and display performance metrics such as **Accuracy, F1-Score, Precision, Recall, ROC AUC**, and a confusion matrix.
   - Automatically compare models to identify the best-performing one.
2. **Model saving:**
   - Save only the best-performing model for future deployment.
3. **Development of a user interface via Streamlit:**
   - Allow users to submit features and receive predictions.
   - Provide the ability to visualize model performance.
4. **Containerization with Docker:**
   - Prepare a containerized application for simple and portable deployment.

3. Deliverables

1. **Functional Python scripts:**
   - Main program for training and evaluating models.
   - Automatic saving of the best model in a `.pkl` file.
2. **Streamlit application:**
   - Interactive user interface for making predictions with the saved model.
3. **Docker image:**
   - Docker container with the application and its dependencies, ready to be deployed locally or in the cloud.
4. **Model comparison report:**
   - A summary table of the performance of the different models (Accuracy, F1-Score, etc.).

4. Project Steps

1. **Data analysis and preparation**
   - Load and clean the data.
   - Separate features (**X**) and labels (**y**).
   - Split the data into training and testing sets.
2. **Model implementation**
   - Create multiple supervised models:
     - **K-Nearest Neighbors (KNN)**
     - **Random Forest**
     - **Support Vector Machine (SVM)**
     - **Logistic Regression**
     - **Naive Bayes**

   - Compare their performance using defined metrics.
3. **Automation of the best model selection**
   - Compute metrics for each model.
   - Automatically identify and save the model with the highest **F1-Score** (or another chosen metric).
4. **Development of a Streamlit interface**
   - Create an interactive interface to:
     - Submit features and obtain predictions.
     - Visualize the overall performance of models (table and confusion matrix).
5. **Containerization with Docker**
   - Prepare a `Dockerfile` to containerize the Streamlit application.
   - Create a Docker image to deploy the application on any environment.
6. **Documentation**
   - Include a report explaining the performance of the models.
   - Write user documentation for running the Streamlit application and Docker container.

5. Success Criteria

1. **Minimum accuracy:**
   - The selected model must achieve an **F1-Score > 85%** on the test data.
2. **User interface:**
   - The Streamlit application must be intuitive and functional.
   - Allow easy data submission and display clear predictions.
3. **Successful containerization:**
   - The application must run seamlessly within a Docker container.
4. **Performance report:**
   - Provide a clear table comparing the performance of the models.

6. Target Audience

- **Data Analysts:** To analyze model performance and automate predictions.
- **Developers:** To quickly integrate predictions into their applications using the selected model.
- **End Users:** To interact with the model through a simple interface.

7. Technologies Used

- **Python** (with libraries: `scikit-learn`, `joblib`, `streamlit`, `matplotlib`, `seaborn`, `pandas`, `numpy`)
- **Docker** (for containerization)
- **Streamlit** (for the user interface)

Final Goal of the Project

To create a complete and portable supervised classification solution that:

- Identifies the best-performing machine learning model for specific data.
- Offers a user-friendly interface for making real-time predictions.
- Provides a containerized application easily deployable in any environment.