Technical Specification Double-side Page

1. **TECHNICAL SCOPE:** Summarise the mock-up devised during the EXPLORE phase: how have you addressed the challenge/Theme Challenges and tackled with its requirements and data. Include a diagram.

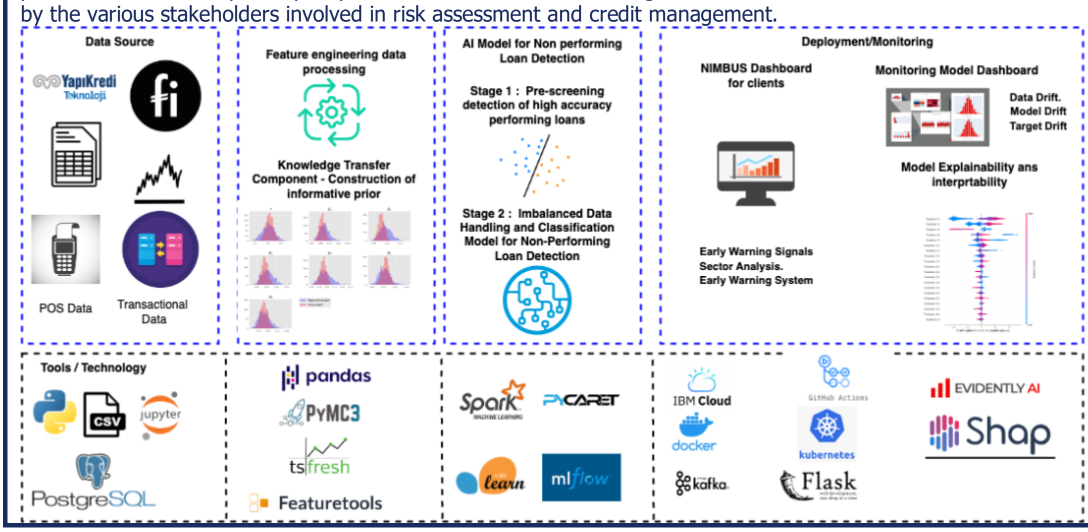
Finclude developed the NIMBUS (Non-performing loans Identification through Machine learning Based Underwriting and Screening) project to tackle the challenge of early detecting Non Performing Loans (NPLs) of large-scale firms. NIMBUS employs a two-stage classification model that combines both frequentist and Bayesian approaches, leveraging in-house financial data, customer behaviour insights, and data from Yapi Kredi Teknoloji. This robust framework provides valuable risk assessment and credit management insights.

The project leverages Bayesian principles, such as the construction of an informative prior, to transfer knowledge gained from past customers with longer data histories to the analysis of new customers. By incorporating this Bayesian approach, NIMBUS enhances the accuracy of NPL detection and enables stakeholders to make more precise risk assessments.

The project delivers a comprehensive dashboard designed to benefit various stakeholders, including risk managers, credit analysts, and decision-makers. The intuitive interface serves as an early warning system, categorising customers into two groups. The first category identifies those likely to experience delinquency in loan payments between 11 and 90 days within the next 6 months, allowing for proactive measures. The second category focuses on customers expected to have delinquency in loan payments over 90 days within the same timeframe, enabling effective risk mitigation.

NIMBUS employs a two-phase approach, starting with the detection of performing loans that are comparatively easier to identify. This initial step serves as a foundation for subsequent analysis, where non-performing loans are detected using techniques tailored for imbalanced datasets. By leveraging the knowledge transferred through the Bayesian approach, particularly for customers with limited data, the model achieves heightened accuracy, empowering stakeholders to anticipate delinquencies and take timely actions.

In addition to NPL detection, NIMBUS provides a monitoring dashboard that allows stakeholders, including risk managers and credit analysts, to track data model performance and identify potential target variable drift. This real-time monitoring ensures the system's ongoing reliability and accuracy. Moreover, the project prioritises the use of interpretable and explainable algorithms, enabling stakeholders to gain insights into the factors driving model predictions. This transparency empowers informed decision-making and effective communication of actions taken by the various stakeholders involved in risk assessment and credit management.


Data Source
YapıKredi Teknoloji
fi
POS Data
Transactional Data
Feature engineering data processing
Knowledge Transfer Component - Construction of informative prior
AI Model for Non performing Loan Detection
Stage 1 : Pre-screening detection of high accuracy performing loans
Stage 2 : Imbalanced Data Handling and Classification Model for Non-Performing Loan Detection
Deployment/Monitoring
NIMBUS Dashboard for clients
Early Warning Signals
Sector Analysis.
Early Warning System
Monitoring Model Dashboard
Data Drift.
Model Drift
Target Drift
Model Explainability ans interprtability
Tools / Technology
CSV
jupyter
PostgreSQL
pandas
PyMC3
tsfresh
Featuretools
Spark MACHINE LEARNING
PYCARET
learn
mlflow
IBM Cloud
docker
kafka
GitHub Actions
kubernetes
Flask
EVIDENTLY AI
Shap

2. **ALGORITHMS, TOOLS AND CONCLUSIONS:** Detail the algorithms and tools identified to accomplish the challenge/Theme Challenges. Show clear understanding of the used REACH dataset/s and addressed challenge/Theme Challenges.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

The NIMBUS project utilises a range of algorithms and tools to accomplish the challenge of NPL detection for new customers of large-scale firms. Here is a list of the algorithms and tools employed:

1. Feature engineering: Tools such as pandas, Featuretools, and tsfresh are utilised for effective feature engineering, particularly for time series data.
2. Informative prior construction: PyMC3, a Python package for Bayesian statistical modelling, and STAN language is used to create experimental models from scratch.
3. Clustering algorithm: Various clustering , algorithms including OPTICS, DBSCAN, K-Means and others. These algorithms are applied to identify similarities among firms, especially in anonymized data. Clustering enables the treatment of clusters as independent groups, facilitating the construction of informative priors tailored to each cluster.
4. Classification algorithms: XGBoost, ExtraTreesClassifier, LogisticRegression, and CatBoost models, fine-tuning them through grid search cross-validation. Pipelines combine the two stages of the model into a unified model.
5. Model management and versioning: The model management and versioning is handled by MLflow, ensuring efficient tracking and comparison of different models for continuous improvement in NPL detection.
6. Model deployment: The NIMBUS project utilises Flask APIs encapsulated as Docker containers deployed on IBM Cloud services. Container orchestration is managed through Kubernetes, while GitHub Actions ensures the CI/CD process.
7. Monitoring dashboard: The Evidently.ai tool is employed to develop a monitoring dashboard, enabling real-time tracking of model and data drifts.

3. **SCALABILITY AND FLEXIBILITY OF THE SOLUTION:** Discuss whether the solution can truly cope with humongous and increasing datasets and how flexible it is to adapt to other related domains and integrate into Data Value Chains (DVC).

The NIMBUS solution is designed with scalability and flexibility in mind, allowing it to effectively handle large and growing datasets. By leveraging distributed computing frameworks like Apache Spark, the solution can efficiently process and analyse large datasets, ensuring its capability to cope with increasing data volumes.In addition to the aforementioned scalability and flexibility, the NIMBUS solution leverages the capabilities of the SparkML library to ensure efficient and scalable machine learning modelling. SparkML provides a robust and distributed framework for training and deploying machine learning models, making it well-suited for handling large-scale datasets within the NIMBUS solution.

Moreover, the NIMBUS solution is designed to be adaptable to other related domains. Its modular architecture and flexible design make it easier to integrate into different Data Value Chains (DVC). The solution's components, such as the feature engineering pipelines, clustering algorithms, and classification models, can be customised and extended to address specific requirements and domain-specific challenges.

Additionally, the use of industry-standard tools and technologies, such as Docker and Kubernetes, enables seamless deployment and scalability across various environments. This flexibility allows the NIMBUS solution to be easily deployed in different infrastructures, whether on-premises or in the cloud, ensuring compatibility and adaptability to different IT landscapes.

4. **DATA GOVERNANCE AND LEGAL COMPLIANCE:** Describe the security level of the proposed solution, i.e. how authentication, authorization policies, encryption or other approaches are used to keep data secure. Explain how will be compliant with the current data legislations concerning security and privacy (e.g. GDPR).

Finclude is a regulated fintech and complies with all relevant privacy and security regulators as described in the guidelines issued by the Central Bank of Ireland and the European Banking Authority. Our operations and processes are monitored by the Central Bank of Ireland, as well as audited by independent third parties and our security is also reviewed by third parties to ensure compliance with the required standards and regulations. We perform an annual test and review on our security policy, access management policy and BCP as well as an annual pen-test on our apps & infrastructure.

5. **QUALITY ASSURANCE AND RISK MANAGEMENT:** Describe the quality process planed for the final product. Technologically, which are the potential risks in all the phases of the project (design of the solution, development, testing, deployment…) and indicate mitigation plans to still fulfil the challenge/Theme Challenges and data provider requirements.

To ensure the final product's quality, a rigorous quality process will be implemented, including extensive testing and validation of the models before deployment. This process will involve testing the models on diverse datasets to verify their accuracy and effectiveness. Regular code reviews and adherence to software development best practices will ensure maintainable and scalable code. Working with anonymized data presents a potential risk, limiting a comprehensive understanding of the processed information. To mitigate this, acquiring the complete dataset is planned to gain deeper insights into the data.The project addresses the challenge of sparse real production data by leveraging a combination of Bayesian and frequentist approaches. This approach improves the accuracy of the models and tackles the challenges posed by unbalanced datasets.

To mitigate the risk of creating a black-box model, fundamental indicators will be enriched using machine learning techniques. The implementation of meta-models will enhance interpretability, allowing for a refined understanding of the models while maintaining transparency and explainability.

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 951981