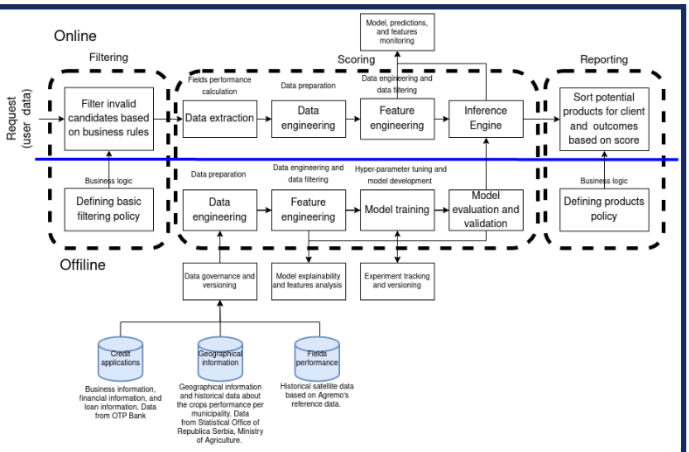
Technical Specification Double-side Page

1. **TECHNICAL SCOPE:** Summarize 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.



The solution of our project will represent a platform that provides an automatic response regarding the creditworthiness of RAH clients that have a considerably larger amount of data and indicators than other clients. The AI and ML models that are to be implemented in the solution are built based on the analysis of the correlation between the credit activity, the quality of repayment of credit products, and data within the scope of business with RAH clients (date of registration, district, the area and types of crops, animal husbandry, the ownership status of the land, work activity, income from agriculture, other incomes, etc.). Besides the credit applications data, the model will process other external data sets. These include the geographical information and historical data about the crops performance per municipality (provided by the Statistical Office of Republic of Serbia, Ministry of Agriculture, and Agro department of OTP bank) and clients fields performance based on satellite imagery (provided by the primer Agremo products).

The solution will automatically filter the invalid candidates, based on the business rules and filtering policy. Online processing of data include Data extraction, Data, Feature, and Inference engineering. Offline processing includes Data and Feature engineering, as well as model training, evaluation, and validation. Offline evaluation will be performed using historical data. Based on a vast amount of described data, the solution will provide an automatized information to final user - the bank, whether to credit the applicant or not. The main output of the solution will be automatic, real time, platform provided response to an RAH client whether the loan is approved or not. This provides the speed and promptness of the response, so banks would save clients and prevent customer churn. The second output of the solution will involve offering bank clients a list of credit products for which they are eligible.

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.

The primary goal of AI and ML in this project is to model and predict the RAH credit loan approval. Second output presents a list of credit products client are eligible for, that bank might offer if necessary. For the solution development, we will use **Python** and its related libraries. In the process of implementation, several models will be trained, assessed, and compared. AI and ML models to be used and evaluated in the process include **Deep Neural Networks, Random Forests, XGBoost, Support Vector Machines**, and similar models. For the different outputs, different models might be required. Our solution will take very seriously the following concepts: understandability, comprehensibility, interpretability, explainability, and transparency. Model-agnostic techniques for post-hoc explainability will be used to address these concepts, such as LIME, ICE, or SHAP. Data versioning together with experiment tracking and versioning will be used during ML cycles aiming to guarantee repeatability, reproducibility & replicability. Random Forests and SVM had proven to have superior performance over other models, so far.

Models and data will be monitored by tracking, measuring, and logging, aiming to quickly identify technical risks, such as data drifts, distributions shifts, or other odd behaviour. If identified, these problems will be attacked in the offline part of our solution. Problems with data will be addressed in the data engineering (preparation) phase. Problems with the model will be addressed during the training and evaluation phase with proper techniques or different approaches.

In the solution implementation phase, we will fine-tune the input parameters, if necessary, based on the information provided by bank experts. To scale-up the models in the future for a large number of credit applicants we will use **PySpark** as the Big Data framework. Other tools to be used in the implementation of the solution include Anonymizer from REACH Toolsuite, Pandas, Numpy, scikit-learn, open source package InterpretML, TensorFlow, PyTorch, MLFLow, TensorFlow Serving, TensorBoard, Flask, Django, FastAPI, OpenSSL. The provided set of the algorithms and tools ensure the speed and promptness of the solution response, which helps the bank to save the clients and prevents customer churn. By considering above mentioned technical aspects, our project aims to deliver a comprehensive and reliable AI model for bank credit scoring, meeting the highest standards of accuracy, performance, interpretability,

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

scalability, security, model maintenance, and ethical practices.

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).

As for the technical scalability and flexibility of our solution, we have considered the future growth and evolving needs of the credit scoring system. Regarding **scalability**, our solution will be designed to handle increasing volumes of data without compromising performance. We will implement scalable infrastructure and optimized algorithms to ensure that the system can process large datasets efficiently. Our multi-step pipeline was composed taking into account a large number of requests. This allows our solution to accommodate a growing customer base and handle a higher number of credit scoring requests seamlessly. Moreover, besides credit clients' data, our data sources include municipality data on fertility and growth of crops, as well as fields performance based on satellite imaginary provided by Agremo, which increases the scalability of our solution. We understand that the financial industry is dynamic, and requirements can change over time, thus our solution will be built with **flexibility** in mind. Therefore, our architecture and design choices will enable easy adaptability to evolving business needs and regulatory updates. One aspect of flexibility is the modular architecture of our solution, that will allow for independent updates and enhancements to different parts of the system, minimizing the impact on other components. It also facilitates the incorporation of new features or functionalities as required. After the successful implementation with OTP, our aim is to make a global solution for RAH credit loans and serve banks worldwide. We aim to employ adaptable and expandable technologies. By harnessing the power of cloud-based infrastructure and leveraging containerization, alongside orchestration technologies such as Docker and Kubernetes, we will enable effortless deployment and scalability. This guarantees the effortless deployment of our solution in diverse environments. Furthermore, the solution will be designed with API approach, that allow for easy integration with existing and new third-party applications.

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).

From a data security perspective, the solution will implement **Fedehr** access, other access restriction control services, and **OpenSSL** once the model is deployed in production. Integration of all the services, different sources of data and the overall process will be addressed with Fedehr support services. To mask data and achieve compliance with data masking policies as GDPR or CCPA, we will use Fedehr Anonymizer and Fedehr training from REACH Toolsuite. In terms of legal compliance, we work closely with legal experts and compliance teams to ensure that our AI model adheres to relevant laws, regulations, and industry standards. From the privacy perspective, the data will be or already is masked at the source in process since this information is not expected to have any benefits in the AI process. For credit applications of small and medium RAH clients, it means it will be anonymized prior sharing and use. In the same way, the data in production will be masked. The Geographical information and historical data about the performance of the crop per municipality is public data which is already anonymized and presented statistically.

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.

Our **quality** assurance includes data pre-processing and cleansing to ensure data accuracy and consistency. We will conduct extensive testing, including unit tests, integration tests, and end-to-end tests, to verify the functionality and performance of the model. We will also perform cross-validation and use appropriate evaluation metrics to assess the model's accuracy and predictive capabilities. So far, model validation has shown over 96% accuracy in classifying clients based on input variables in the model. Besides, we will also rely on OTP data provider expertise, and, in line with agile principles, we will have regular meetings with OTP and update the solution according to their feedback.

The **risks** include data shift or odd behaviours. Data on production might differ from the training/test data or a specific event might cause data shifts. Model performance, predictions, and features monitoring can identify these shifts, and continual learning mitigates these issues. Unstructured data or poor data might lead to customer segmentation or even biased models, when the data is not structured in the same way over time and across the different organisation units. Monitoring tools will help us to identify these problems and the previously mentioned tools for model interpretability and transparency will help us to understand the source of these problems. Such information is essential to take the proper actions aiming to mitigate them, such as increase data diversity and representation, model regularisation or fairness techniques, among others. Besides that, it is important to highlight that we will determine clear guidelines for data input for clients and we will have an automated data collection process, which should help to reduce problems with data format. Lack of transparency shown in poor model interpretability and explainability might lead to lack of confidence. This will be addressed by using the SHAP algorithm to provide a clear insight into the model components. Our model will be designed from the start to be transparent and explainable,

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

enabling auditors, regulators, and stakeholders to understand the factors contributing to credit scoring decisions.

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