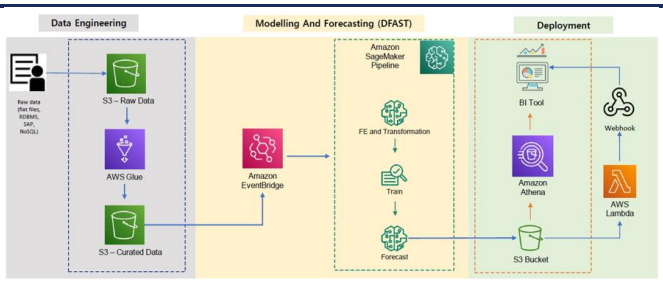
Technical Specification Double-side Page

Name of the Company: **Amplify Analytix**
Challenge: Forecast of production needs
TRACK: 3

1. **TECHNICAL SCOPE:** Summarize the solution developed during the EXPERIMENT phase: how have you finally addressed the challenge/Theme Challenges and tackled with its requirements and data. Include a diagram. Mention the different shared industrial and open data used.

Amplify's DFAST solution addresses the problem of forecasting for production planning for a multinational food packaging manufacturer, Tetra Pak. It provides highly accurate forecasts, can scale to 100s of individual data series analyses simultaneously, ingests and analyses large data sets quickly and incorporates use of external data. Our innovative use of model competition – at the level of the individual data series - ensures unparalleled flexibility and competitive advantage: rather than compete with giants like FB Prophet, we harness the capabilities of 12 such best of breed models to provide the most accurate result.



Improvements accomplished during Experiment stage: (1) Ensuring secure DVC through building authentication step plus user log-in module, (2) Streamlining data input and data prep for easier use and adding UI functionalities around selecting the dimension and level of the forecast, identifying null values and imputation by two methods, tackling outliers by smoothening techniques, (4) Ensuring maximum accuracy and flexibility by enabling competition among 12 classical to state-of-art models, incorporating external factors to improve forecast quality and allow analysis of non-stationary data, (5) Forecasting features around accuracy evaluation criteria (MAPE), faster selection performance, scenario-based business oriented forecasting with exogenous variables (benefit beyond the scope of the challenge, allowing experts to construct forecasting scenarios based on domain knowledge), and (6) Closing the loop on secure DVC by secure pushing output to AWS S3.

Data used: For the live demo and development, we used Tetra Pak's internal sales data on finished goods at various locations. External (exogenous) data: (1) FAO Dairy Price Index that is an important open-source dairy industry index in the value chain of Tetra Pak, and (2) Consumer Inflation data that drives demand across many industries. For the tool link provided below, for data privacy reasons, we use Kaggle open-source data with similar structure as our client's.

2. **ALGORITHMS, TOOLS AND CONCLUSIONS:** Detail the algorithms and tools finally selected to accomplish the challenge/Theme Challenges. Summarize the main results that you have obtained during the EXPERIMENT phase: data, insights, conclusions and the main contributions to solve the challenge/Theme Challenges.

We successfully incorporated 12 Econometric and ML models that can handle a large variety of use cases, more specifically: (1) Univariate models that decode trend-seasonality compositions efficiently SNAIVE, ETS (multiple variations), THETA, THIEF; (2) Multivariate models that leverage macro-economic and industry/value-chain-specific exogenous variables along with historicity, addressing the challenges of small and/or non-stationary series: STLM, ARIMAX, Prophet, NNETAR, MARS, Random Forest, SVM, LR. **Results:** Running the analysis on 10 client series takes 2mins, with average accuracy of 85%. The model competition results in 5 different models being selected, demonstrating the importance of having multiple models competing to achieve the lowest MAPE. Running on 65 client series takes 14mins, with avg accuracy 75%, and all 12 models available are selected at least once.

Main innovation: Incorporating a large number of state of art models in competition allows to perform individual analysis and select the best model for each series, while also not costing much in terms of time. **Data challenges:** DP data can become less smooth - Intermittent, Erratic and Lumpy – when going to lower aggregations like day, product sub-category, location, while typical time series models assume smooth data. Including exogenous variables may help with such irregular patterns (suggested by mentor Miguel Bravo). We would like to research this new challenge because we see commercial potential for solving it.

3. **SCALABILITY AND FLEXIBILITY OF THE SOLUTION:** Explain how the solution copes with the challenge/Theme Challenges requirements and how can it be adapted to other similar problems. What work is still pending to create a real/stable product if any? What TRL level is it in? Is it a DVC?

Forecasting is a universal problem at the heart of various business processes – financial, production, procurement, supply, staff planning and optimization. Our solution aims to tackle the forecasting problem with maximum adaptability by allowing model selection across a verity of methods, and by allowing the use of external variables that bring in information specific to that industry and to that Data Value Chain. On the data architecture side, to ensure scalable performance across many clients with large data sets, we containerized the deployment. We achieved TRL 7, demonstrating functional maturity as our product has transitioned from demo stages to handling client data in production. Adherence to a structured and secure Data Value Chain (DVC) process to maximize data value has been ensured (more details below). Solution can handle large datasets by using Datalake (AWS S3) for raw and processed data storage. Solution can be integrated into multi-stakeholder DVC by sourcing data via secure API from private upstream or downstream entities – plastics producers, supermarkets, dairy producers, etc – to forecast demand or costs for product packaging more accurately. From the agreed backlog, a feature that allows aggregating data (ex, weekly to monthly) is still pending. This unlocks ability to forecast at different frequencies. Pending developments (not in backlog): On modelling side, include introducing lagged feature for ML forecasting; feature for multiple exogenous variables; Integration of models for sparse time series forecasting; Grid Search hyperparameter tuning for ML models. On scalability side, next step is to create a micro-service based architecture. This could be achieved by creating separate images/modules for each task (modules for data transformation, training, forecasting, etc.). Further each task could be divided to sub-tasks, for example model training could have sub-tasks defined by a data series-algorithm, which could be run on separate pods in Kubernetes.(discussed with mentor Adrian Iacovelli)

4. **DATA GOVERNANCE AND LEGAL COMPLIANCE:** Describe the security level of the solution, i.e., how authentication, authorization policies, encryption or other approaches are used to keep data secure. Explain how the solution is compliant with the current data legislations concerning security and privacy (e.g. GDPR). Describe in a convincing way how your solution realises a secure DVC, e.g., through usage of specific tools.

Security & Storage: (1) AWS S3 Buckets: We prioritize data integrity by utilizing AWS's secure S3 buckets. Each user benefits from dedicated storage space, ensuring data isolation and prevention of inadvertent access. (2) Communication Protocols: Our system's communications are fortified with TLS/SSL protocols, safeguarding data transfer against potential interceptions or breaches. (3) Monitoring & Restrictions: We host the app within a specialized AWS account. This account is under constant surveillance to detect and respond to any anomalous activities.
Privacy & Access: (1) Data Anonymization: We tested out and applied the "Anonymiser" tool from the REACH toolbox to the DP data to ensure there's no unintentional leakage of personal data. While with Tetra Pak we use non-personal data, this can be a useful capability for other domains, or if the user would like to anonymize location, product or any other dimension of the data. (2) Session Management: Once a user session concludes, all related data is irrevocably purged.
Compliance: (1) ISO Certification: Upholding global standards, our company proudly holds the ISO/IEC 27001:2013 certification (Certification No: 510/19). (2) GDPR Alignment: We are steadfast in our adherence to the GDPR guidelines.

5. **QUALITY ASSURANCE AND RISK MANAGEMENT:** Describe the quality process followed for the final product. Technologically, which problems have you encountered and how you have solved them, and any processes followed that guarantee that the solution fulfils the challenge/Theme Challenges and data provider requirements

Quality Assurance Process: Our quality assurance approach stems from an agile QA methodology. This framework, Microsoft's TDSP and our hands-on experience across various economic sectors, ensures that our solution remains top tier in both functionality and performance.
Challenges & Their Resolutions:
Design Challenge/ Solution: Our initial aggregated code structure posed a challenge as it rendered modifications and updates cumbersome. We transitioned to a modular code design, which took up a long time to develop. This expedited customization processes and also enhanced debugging capabilities/ quick development.
Technical Development Challenge/ Solution: Creating enhanced forecasts based on exogenous data requires that we have future values of said exogenous data, which is challenging. After research/brainstorming, we innovated a mechanism for scenario-based forecasting for future periods based two methods, extrapolation and user-provided manual input based on domain insights.
Testing Challenge/ Solution: The reliability of our computational engine and its forecasting models was of paramount importance. We adopted an exhaustive 'stress testing' regimen. This involved challenging our system with diverse datasets sourced publicly, ensuring that our models can withstand varying lengths of time series and diverse data formats.
Deployment & Scalability Challenge/ Solution: Given the variability in business needs and datasets, our solution had to be both versatile and adaptable. We architected a streamlined and secure DVC which integrates AWS S3 for data storage, AWS Sagemaker for tailored ETL processes, and finally culminates in our main application, DFAST, hosted securely on an EC2 instance. This structured approach ensures that any customizations or updates are largely restricted to the Sagemaker ETL pipeline, simplifying deployment and adaptations.

Annex 1. Means for accessing the MVP

Please, indicate in 1 page maximum the means for accessing the MVP for a potential customer (login information, website address, link to a demo video or whatever means are needed to check that the MVP exists and works).

Means to assess MVP for potential customer

Mode of assess – through web browser (recommended Google Chrome/ Microsoft Edge in normal mode)

Link - Amplify DFAST (amplifydevops.com)

User id – *demouser1*

Password – *demopass1*

Means to assess data for testing MVP

Link to demo datasets – click here

(Please download the csv file(s) available in the above location and upload in DFAST through "Browse- Upload"

Means to assess Demo Usage of the MVP

Link to demo video

Link to demo doc

In case of any issue with accessing MVP and/ or data set or help docs, please contact us at: *milena.petrova@amplifyanalytix.com*