KGI 04 CASE 1 MACHINE LEARNING CANVAS

Designed for: Microsoft KGI Hackathon | Designed by: KGI 04 | Date: 26/04/2025 | Iteration: 5

PREDICTION TASK

What is the type of task?
Which entity are predictions made on?
What are the possible outcomes to predict?
When are outcomes observed?

*The task is a **regression task** to predict **reservoir fluid production volumes**. The predictions are made on **well samples with associated petrophysical and geophysical measurements**. The possible outcomes to predict are **continuous production volume values** (fluid production quantities). The outcomes are observed after **well development and production** has begun; historical data is used for training.*

DECISIONS

How are predictions turned into actionable recommendations or decisions for the end-user? (Mention parameters of the process / application for this.)

*Predictions help **reservoir engineers optimize drilling locations, plan enhanced recovery strategies, forecast productivity in undrilled locations**, and **allocate resources efficiently**. The parameters are **well-planning parameters**, which include resource allocation, intervention strategies, and production optimization techniques.*

VALUE PROPOSITION

Who is the end beneficiary, and what specific pain points are addressed?
How will the ML solution integrate with their workflow, and through which user interfaces?

*- **Energy companies and reservoir engineers** benefit by gaining **more accurate** production forecasts, **reducing uncertainty** in reservoir management, **minimizing drilling costs**, and **optimizing resource** allocation. The solution **addresses critical pain points** including unpredictable production decline, suboptimal well placement, and inefficient capital expenditure*
*- The model can be deployed **as a decision-support tool** for engineers to predict production based on well log data, enabling proactive rather than reactive management. It **enhances existing workflows** by providing data-driven insights that complement traditional reservoir engineering methods, creating a hybrid approach that leverages both domain expertise and advanced analytics.*

*- **Interactive web-based dashboard** with visualization capabilities for scenario testing and sensitivity analysis; **API integration** with existing reservoir management software systems; **mobile applications for field engineers** to access predictions on-site*

DATA COLLECTION

How is the initial set of entities and outcomes sourced (e.g., database extracts, API pulls, manual labeling)?
What strategies are in place to update data continuously while controlling cost and maintaining freshness?

*- **Historical well data** including **petrophysical** measurements, **geophysical** properties, and corresponding **production volumes**.*

*- **Regular updates** from new well logs and production data; **IoT sensors** for **real-time measurements** of pressure, temperature, and flow rates; automated data pipelines for continuous integration of IoT data streams while filtering for quality control*

DATA SOURCES

Where can we get data on entities and observed outcomes? (Mention internal and external database tables or API methods.)

*- **Internal company databases** containing well log measurements and production histories; **IoT sensor** networks deployed in wells and surface facilities; possibly **external databases** of regional reservoir characteristics*

*- **Direct database extraction** of well measurements; **real-time IoT data streams** from downhole sensors (pressure, temperature, flow rate); **production volumes** from field operations; **SCADA systems** for surface facilities; and **laboratory analysis** of core samples.*

IMPACT SIMULATION

What are the cost/gain values for (in)correct decisions?
Which data is used to simulate pre-deployment impact?
What are the criteria for deployment?
Are there fairness constraints?

*- **Correct predictions** lead to **optimized resource allocation**, **reduced drilling costs**, and **maximized production**; **incorrect predictions** may result in **suboptimal** well placement or resource **misallocation**.*
*- **Test dataset (20% of data)** and **cross validation** is used to **evaluate model** performance before deployment.*

*- R^2 score above acceptable threshold (current best: XGBoost with $R^2 \approx 0.99$), **RMSE minimization**, and **realistic residual distribution**.*
*- Model should **perform consistently** across different reservoir types and geological formations*

MAKING PREDICTIONS

Are predictions made in batch or in real time?
How frequently?
How much time is available for this (including featurization and decisions)?
Which computational resources are used?

*- **Batch prediction** for planning purposes and potentially **real-time** for **active drilling** operations.*

*- **Weekly or monthly** for overall reservoir management; possibly more frequently during active drilling campaigns*

*- Predictions can be generated within **seconds to minutes** after input data is available. **Lightweight computing resources**, the final pipeline includes standardization and prediction steps.*

BUILDING MODELS

How many models are needed in production?
When should they be updated?
How much time is available for this (including featurization and analysis)?
Which computation resources are used?

*- **One final model (XGBoost** with optimized hyperparameters) is deployed, with other models (Linear Regression, Random Forest) used for **comparison** and **benchmarking**.*
*- Models should be updated when **new significant data** becomes available or when **model drift** is detected through monitoring.*
*- Initial model development takes **several hours** for **feature engineering, training, and hyperparameter optimization**.*
*- **Standard computing resources** for training; hyperparameter optimization uses **parallel processing** (n_jobs=-1)*

FEATURES

What representations are used for entities at prediction time?
What aggregations or transformations are applied to raw data sources?

*The representations used for prediction include **petrophysical and geophysical properties** are consisting of **Porosity (Por)**, **Log Permeability (LogPerm)**, **Acoustic Impedance (AI)**, **Brittleness Index (Brittle)**, **Total Organic Carbon (TOC)**, and **Vitrinite Reflectance (VR)**. Transformations applied to the raw data sources include **outlier detection and handling** using the **Interquartile method** (IQR method), **standardization** of features using **Standard Scaler**, **logarithmic transformation** for permeability.*

MONITORING

Which metrics and KPIs are used to track the ML solution's impact once deployed, both for end-users and for the business?
How often should they be reviewed?

*- Model performance metrics (**RMSE**, **R^2**, **MAE**), prediction accuracy over time, business impact metrics (**resource efficiency, production optimization percentage**).*
*- **Monthly** for technical performance; **quarterly** for business impact assessment.*