KGI 04 CASE 2 MACHINE LEARNING CANVAS

Designed for: Microsoft KGI Hackathon Designed by: KGI 04 Date: 26/04/2025 Iteration: 12.

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?

- Regression task to predict sonic travel time (DT) values in well logs.
- Well measurements at specific depths within boreholes
- Continuous DT values (sonic travel time in microseconds per foot)
- After well logging operations are completed, but DT logs may be missing or unreliable in certain sections

DECISIONS

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

- Predicted DT values enable more complete formation evaluation, allowing geologists and engineers to make better informed decisions about porosity calculations, lithology identification, and seismic-to-well ties.
- Params: Formation properties interpretation, velocity model creation, synthetic seismogram generation, and reservoir characterization parameters

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?

- Geologists, geophysicists, and reservoir engineers benefit by gaining access to complete DT logs even when original measurements are missing or corrupted. This addresses the critical pain points of data gaps that compromise interpretation quality, formation evaluation accuracy, and seismic correlations.
- The model seamlessly integrates into the existing petrophysical interpretation workflow, automatically filling gaps in sonic logs while maintaining geologically reasonable values. It enhances rather than disrupts established processes by providing high-quality inputs to downstream applications.

- Interactive log visualization tools with predicted vs. actual log overlays; integration with industry-standard petrophysical software platforms; API connections to reservoir modeling software; custom dashboards for real-time log quality assessment and enhancement.

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?

- High-quality well log data from the Volve field, containing comprehensive measurements across multiple wells with varying depths and geological formations.
- Automated integration with real-time logging-while-drilling (LWD) systems; periodic retraining with new well data; quality control workflows to identify and flag anomalous predictions; data validation protocols to ensure geological consistency.

DATA SOURCES

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

- Public Volve dataset from Equinor; internal company well log databases; data exchanges with partner operators; LWD and wireline logging service providers.
- Direct import of LAS or DLIS log files; database queries from corporate data repositories; API integration with cloud-based log storage systems; extraction from geological interpretation software.

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?

- Accurate DT predictions enable proper formation evaluation, which directly impacts drilling target selection and completion design worth millions of dollars; errors could lead to missed pay zones or ineffective completions.
- Holdout well data (testing on a completely different well than training data) to validate performance in realistic scenarios
- Low RMSE (< 10 μs/ft), geologically consistent predictions, and smooth log curves verified by domain experts
- Model must perform well across different lithologies and depth ranges; no systematic bias in predictions related to specific formation types.

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?

- Primarily batch processing for entire well sections; potential real-time application during logging operations.
- On-demand for completed wells; potentially continuous updates during active drilling operations.
- Predictions must be available within minutes after log data acquisition to support timely decision-making during operations.
- Lightweight inference pipeline (scaler + trained model) that can run on standard workstations or be deployed as a cloud service.

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 (selected from Random Forest, XGBoost, Neural Network, or Linear Regression based on performance) is deployed after comparative validation.
- Models should be retrained when substantial new well data becomes available or when geological conditions significantly differ from the training dataset.
- Initial model training requires several hours for cross-validation and hyperparameter optimization; subsequent updates can be performed overnight to incorporate new data.
- Standard compute resources for model training; hyperparameter tuning with Optuna using parallel processing; Azure ML infrastructure for deployment

FEATURES

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

- Raw well log measurements (NPHI, RHOB, GR, RT, PEF, CALI) and engineered features combining these logs (NPHI_RHOB, GR_NPHI, PEF_RHOB, RT_LOG).
- Log normalization using Robust Scaler; outlier detection and handling via IQR method; domain-specific and advanced imputations including K-Nearest Neighbors and Random Forest imputations; creation of physics informed feature combinations; logarithmic transformation of resistivity values.

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?

- Technical metrics: prediction accuracy (RMSE, R^2, MAE), prediction vs. actual curves, residual distributions; Business KPIs: reduction in log acquisition costs, improvement in reservoir characterization accuracy, increased confidence in seismic interpretations

- Technical metrics: reviewed immediately after each prediction run; model performance: after each new well; business impact: quarterly evaluation of value created.