Final Report

Predictive Analytics for Oil Well Failures: A Machine Learning Approach

Titilola Oduwole and Jahir Gutierrez

SYST1059A – Capstone Project-BAA 2023-2024

Hend Negm, Russell Allen and Israel Martinez Perez

Executive Summary

This executive summary provides a high-level overview of the research we conducted to predict producer oil well failures using machine learning techniques. This project focuses on addressing significant challenges in the oil and gas industry, specifically improving the prediction and prevention of oil well failures to enhance operational efficiency and economic performance.

- **Problem Definition / Research Objectives**
  The oil and gas industry continually seeks innovative solutions to enhance operational efficiency and ensure the reliability of its production systems. One critical challenge is the prediction and prevention of oil well failures, which significantly impact production rates, safety, and environmental sustainability. In 2023 alone, the studied oil field experienced 208 failure events, each costing an average of USD 130,000 in intervention costs, excluding deferred production costs. The primary objectives of this project are to develop a robust machine learning (ML) model to predict failures in oil wells before they occur, with a focus on early detection to facilitate timely maintenance, reduce oil well downtime, and improve failure prediction accuracy to mitigate operational and economic impacts. (Titilola Oduwole, 2024)

- **Research methodology / Research questions**
  This research utilized the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, a structured approach encompassing six phases: Business Understanding (the business area being the oil and gas industry, focus on oil wells and the downhole equipment used in production), Data Understanding (database files were Access files and Excel files, converting all files to CSV files, recognising data entry were in Spanish and which files were sourced from the fields, laboratory, etc.), Data Preparation (Dealing with missing data using median imputation, removing duplicates, anonymization of sensitive information, removal of irrelevant columns and merging of dataset into one final file) Modeling (Logistic Regression, Decision Tree, XGBoost, Random Forest and Support

Vector Machine), Evaluation (using the metrics Accuracy and F1 score), and Deployment. The data, covering a period from 2014 to 2023, was collected from various sources, including historical production information per day, fluid level monitoring, laboratory results, and wellhead inspections all gotten from a technical advisor from a Latin-American oil company. The project focused on analyzing historical data to identify failure patterns and develop a predictive model.

Key research questions included:

- What factors contribute to the prediction of oil well failures, particularly focusing on downhole artificial lift systems?
- How can early detection facilitate timely maintenance actions?
- Are there significant trends or patterns in the data that can inform and prevent potential downtimes? (Titilola Oduwole, 2024)

- **Answering research questions and other findings**

The modeling phase involved using five (5) machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, XGBoost and Support Vector Machine, with XGBoost emerging as the best performer with an accuracy of 95.42%. The model successfully predicted the well intervention type and timing for downhole failure events in oil producer wells, offering valuable insights for maintenance and operational decision-making. Significant features identified included Total Fluid rate (bls), Fluid Level (ft), BSW (%), Chlorides (ppm), Run Life (days), Pump Intake (ft), Artificial Lift System, THP (psi), CHP (psi). (Jahir Gutierrez, Modelling Analysis Report - Predictive Analytics for Oil Well Failures: A Machine Learning Approach, 2024)

By implementing the developed ML model, the organization can proactively manage well maintenance, reduce intervention economical impacts, and minimize deferred production. This predictive capability not only enhances operational efficiency but also supports environmental sustainability by preventing incidents that could lead to spills or other ecological damage.

This executive summary encapsulates the essence of the research, highlighting the critical problem addressed, the structured methodology applied, and the significant findings that ensures the achievement of the project's objectives of providing the industry with data-driven recommendations in the prediction of oil well failures which is a big milestone for the oil and gas industry.

Discussion

Findings and Insights

The primary goal of this project was to develop a robust Machine Learning model to predict oil well failures to enhance decision-making and operational efficiency. Through comprehensive data analysis, several key findings were identified:

1. **Predictive Accuracy:** The XGBoost model achieved the highest accuracy at 95.42%, significantly outperforming other models such as Decision Tree and Logistic Regression. This model's performance was validated through rigorous testing, confirming its reliability in predicting oil well failures. Results are shown below: (Jahir Guiterrez, 2024)

   Accuracy: 0.9542
   Precision: 0.9556
   Recall: 0.9527
   F1 Score: 0.9541

2. **Feature Importance:** The analysis revealed that certain features had a significant impact on the model's predictions in this order of importance Total Fluid Rate (bls), Fluid Level (ft), BSW (%), Chlorides (ppm), Run Life (days), Pump Intake (ft), Artificial Lift System, THP (psi) and CHP (psi). SelectKBest() method, using chi-square was used to identify these features. See the output below: (Titilola Oduwole J. G., 2024)

   Feature ranking:
   Total fluid rate (bls) 1
   Fluid Level (ft) 1
   BSW (%) 1
   Chlorides (ppm) 1

Run Life (days) 1

Pump Intake (ft) 2

Artificial Lift System 3

THP (psi) 4

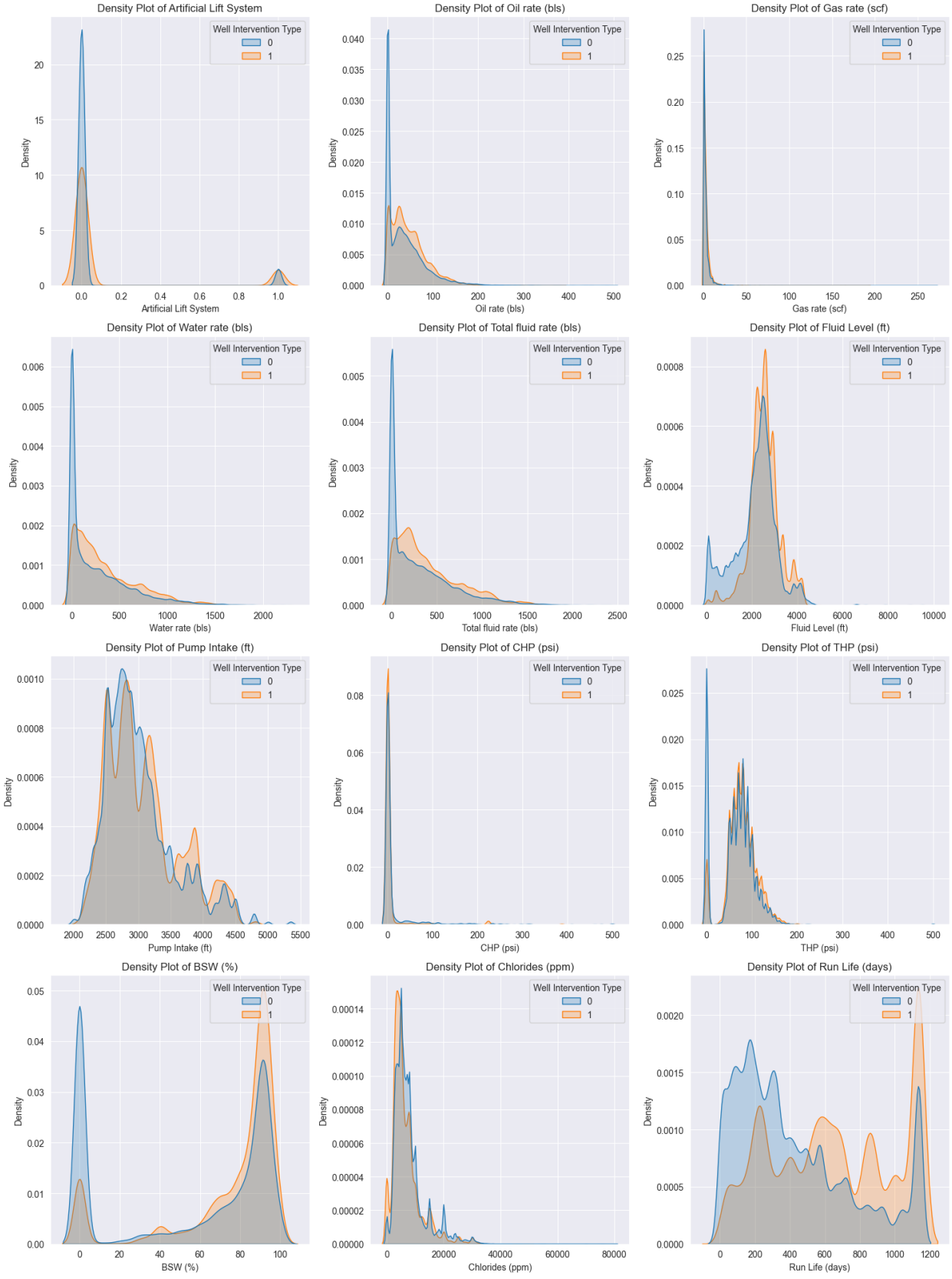
CHP (psi) 5

3. **Correlation Insights:** Fluid level (ft) and Total Fluid rate (ft) showed moderately positive correlation with Run Life suggesting that higher fluid levels are associated with longer run life and higher total fluid rate might be associated with longer run life. Notably, 'Total Fluid Rate' showed a strong positive correlation with BSW (%)' (0.67), indicating that higher fluid rates are associated with higher basic sediment and water percentages due to the waterflooding recovery method used in the field. (Jahir Gutierrez, Modeling Analysis_Decision Tree, 2024)

4. **Density Plot Analysis:** The density plots reveal significant differences in various operational parameters and intervention categories between the two well intervention types (0 – No Failure, 1 – Failure). Key insights include:

- **Intervention Type 1 (Failure)** tends to have longer run lives (for ESP), higher BSW percentages around 80%, and more variability in pump intake and fluid levels. These wells also show lower chloride concentrations, indicating a mature waterflooding process.
- **Intervention Type 0** shows a bimodal distribution in BSW percentages (0% and 100%), more concentrated peaks at lower fluid and water rates, and generally lower oil production rates. These wells exhibit high density at lower THP, CHP, and gas rates, with more variability in pump intake depths and run lives.

  See graphs below:


Density Plot of Artificial Lift System
Artificial Lift System
Density Plot of Oil rate (bls)
Oil rate (bls)
Density Plot of Gas rate (scf)
Gas rate (scf)
Density Plot of Water rate (bls)
Water rate (bls)
Density Plot of Total fluid rate (bls)
Total fluid rate (bls)
Density Plot of Fluid Level (ft)
Fluid Level (ft)
Density Plot of Pump Intake (ft)
Pump Intake (ft)
Density Plot of CHP (psi)
CHP (psi)
Density Plot of THP (psi)
THP (psi)
Density Plot of BSW (%)
BSW (%)
Density Plot of Chlorides (ppm)
Chlorides (ppm)
Density Plot of Run Life (days)
Run Life (days)
Well Intervention Type
0
1
Density

Comparison with Initial Hypotheses

Initially, it was hypothesized that the deployment of advanced machine learning models would significantly enhance the accuracy of predicting oil well failures. The results confirmed this hypothesis, with the XGBoost model performing exceptionally well with an Accuracy of 95.42%. However, there were some unexpected findings:

1. Model Performance: Although the XGBoost and Decision Tree models showed high accuracy, the Logistic Regression model exhibited lower performance in predicting oil well failures due to complex relationships between the features of the dataset.

2. Correlation Assumptions: Statistical Analysis led to the initial assumptions of many numerical variables showing weak or no correlation with 'Run Life (days), leading to the consideration of Well Intervention Type as a more suitable target variable and the creation of the Total Fluid Rate variable. This indicates that the longevity of well operations is influenced by more complex, multifactorial interactions rather than straightforward linear relationships**.** (Titilola Oduwole J. G., Statistical Analysis and Data Visualization, 2024)

3. Unexpected Patterns: The data revealed some surprising patterns, such as the high density of 'Total Fluid Rate' at lower production rates and the significant presence of outliers in 'BSW (%)' and 'FLU_LE.' These outliers are explained by the waterflooding impact (for BSW (%) and by the lack of log type classification (Dynamic / Static Measure) needed for FLU_LE.

4. **Unexpected Insight:** The correlation between Well Intervention Type, Run Life (days) and Artificial Lift Systems showed that ‘ESP’ Artificial Lift System had longer Run Life than PCP and ROD Artificial Lift Systems. See graph below: (Titilola Oduwole J. G., Statistical Analysis and Data Visualization, 2024)

Broader Context and Practical Relevance

The findings of this project have significant implications for the oil and gas industry, particularly in enhancing operational efficiency and reducing economic losses due to unexpected equipment failures. By implementing the predictive model developed, companies can proactively manage well maintenance, thereby reducing intervention costs and minimizing production delays. This not only improves economic performance but also contributes to environmental sustainability by preventing potential spillages and other ecological impacts.

Furthermore, the project underscores the importance of continuous model refinement and the inclusion of additional variables to capture the multifaceted nature of well performance. The insights gained from this analysis can inform future research and development in predictive maintenance, not only within the oil and gas sector but also in other industries where equipment reliability is crucial.

In conclusion, this project has successfully demonstrated the application of machine learning in predicting oil well failures, providing actionable insights that can drive better decision-making and operational efficiencies in the oil and gas industry. The findings

highlight the value of advanced analytics in addressing complex industrial challenges and pave the way for future innovations in predictive maintenance.

Limitations and Challenges

In this section, we explore the various challenges and limitations encountered during the project, along with their impact and the strategies employed to mitigate them**.**

- **Challenges faced:**
  - **Data-related**
    - **Data Quality and Inconsistency:**
      - **Language Inconsistency:** The dataset contained numerous entries in Spanish, posing comprehension and integration challenges. A member of the team speaks and reads Spanish fluently and ensured translation to English was done accurately.
      - **Variable Format Inconsistency:** The "Well Name" and "Date" formats varied significantly across tables, complicating data merging and aggregation.
      - **Missing Data:** Several crucial variables had substantial missing entries, such as 'TYPE_SARTA' with 10.25% missing data and 'SLA_SARTA' with 48.28% missing data.
      - **Unused Data:** Some columns like Rift Subsytem, Component Primary, Rift Cause General were not included in our analysis due to the fact that there are mainly only just recently included in data collection by the oil company.
    - **Duplicate Records:** The presence of duplicate records in variables like 'Run time (days)' and 'BSW (%)' required careful evaluation and cleaning, such as manually calculating for Run Life (days) and BSW (%) for the final dataset.

- **Data Anonymization:**
  - To address confidentiality concerns, sensitive information such as well identifiers and geographic locations were anonymized, potentially limiting the context for analysis.

- **Methodological-related**
  - **Model Performance Variability:**
    - While the XGBoost and the Decision Tree model demonstrated high accuracy (95.42% and 94.45% respectively), other models like Logistic Regression performed moderately well with an Accuracy result of 66.75%, probably due to the complexity of the relationships between variables. Random Forest and Support Vector Machine did not output any results despite many runs due to the size of the dataset (our dataset was for a record of 10 years and probably needed a more robust computational resources than what we had)
  - **Correlation and Multicollinearity:**
    - Weak or no correlation was observed between several numerical variables and 'Run Life (days)', indicating that well performance is influenced by complex, multifactorial interactions.

- **Resource-related (time, student expertise, software/compute limitations)**
  - **Time and Expertise:**
    - The extensive data cleaning and preparation process, combined with the need for domain-specific expertise, required significant time and effort which led to several necessary modification of the dataset to ensure its suitability

for Machine Learning processes. This is seen in changes made to the Well Intervention Type, Artificial Lift Systems to binary variable and in the creation of new features like Total Fluid Rate and manual calculations of Run Life (days) and BSW (%).

- Limited computational resources constrained the ability to explore a broader range of machine learning models and hyperparameter optimizations. At the time of submitting this report, we were yet to get metric outputs from Random Forest and Support Vector Machine models and so could not make a comparison among five (5) algorithms as originally preferred but could only do so among three (3).

- **Limitations of the findings**
  - **Generalizability**
    - The model was trained on a dataset from a specific Latin - American oil company, which may limit its generalizability to other geographical regions or companies with different operational practices and environmental conditions.
  - **Model Accuracy and Uncertainty:**
    - Although the XGBoost model achieved high accuracy, the learning curve showed that the model could benefit from additional data, thereby improving performance, though the rate of improvement may be slower.
  - **External Factors:**
    - External factors such as regulatory changes, market conditions, and technological advancements were not accounted for in the model, potentially impacting its predictive accuracy and relevance over time.

- **Impact of the limitations and challenges**

  The challenges and limitations impacted the project's outcomes in several ways:

  - **Accuracy and Reliability:** A very high data quality and model performance ensures the reliability and robustness of the predictive insights.
  - **Operational Relevance:** Anonymization and limited contextual data may reduce the practical applicability of the findings for specific operational decisions.
  - **Generalizability:** The specific dataset and contextual constraints may limit the broader applicability of the model to different settings or oil industries in different regions.

- **Mitigation strategies**

  - **Improving Data Quality:**
    - Implementing rigorous data cleaning and standardization processes to address inconsistencies and missing values.
    - Applying data imputation techniques for categorical variables with missing entries, and data exclusion for variables with minimal missing data.
  - **Enhancing Model Performance:**
    - Utilizing advanced feature engineering techniques, such as Chi-Squared tests and Recursive Feature Elimination, to identify the most significant features.
    - Addressing data imbalance using techniques like SMOTE (Synthetic Minority Over-sampling Technique) to ensure balanced class distributions in the training data.
    - Manually calculating BSW (%) values and Total Fluid Rate to ensure accuracy and consistency in the dataset.

        - Recording the distinct values in Well Intervention Type as two (2) categories for better adaptability in Machine Learning models.

    - **Leveraging Multiple Models:**

        - Combining different processes like feature selection using selectKBest and RFE, performing PCA and resampling using SMOTE in our different machine learning models, to improve overall model performance and evaluation results.

        - Exploring at least three (3) modelling algorithms and conducting learning curve evaluation on our models to ensure there is no underfitting or overfitting in our models.

    - **Stakeholder Engagement:**

        - Regular communication with our instructors and review of literature on similar projects to validate decisions taken and to ensure the model aligns with operational realities and expectations.

By acknowledging and addressing these limitations and challenges, the project aims to provide a credible and transparent analysis, contributing valuable insights while recognizing the scope for future improvements and broader applicability.

Suggestions

- **Key insights recap**

  The analysis of data related to oil well failures has provided several key insights:

  1. **Predictive Model Performance**: The XGBoost model achieved the highest accuracy (95.42%), highlighting its potential for accurately predicting oil well failures. The Decision Tree model also showed good performance with an accuracy of 94.45%, providing a balance between accuracy and interpretability.
  2. **Important Features**: Key features influencing the model's predictions included Total fluid rate (bls), Fluid level (ft), BSW (%), Chlorides (ppm) and

Run Life (days) – having the ranking of 1 feature selection, Pump Intake (ft) – having the ranking of 2, Artificial Lift System – having the ranking of 3, THP (psi) – having the ranking of 4 and CHP (psi) having the ranking of 5. These features were identified through various feature engineering processes and were found to significantly impact well performance.

3. **Correlation Analysis**: The correlation between Well Intervention Type, Run Life and Artificial Lift System showed that ESP has longer Run Life. Moderate positive correlation between Run Life (days) and Fluid Level (ft) – 0.34, suggests that higher fluid levels are associated with longer run life. The correlation between Run Life (days) and Total fluid rate (lbs) – 0.20 also suggests that higher fluid rates might be associated with longer run life.

- **Recommendations for action including a priority ranking**

Based on these insights, the following recommendations are made, prioritized by their potential impact:

1. **Execute more advanced machine learning Algorithms** (High Priority):
    - Given that Random Forest and Support Vector Machine models did not execute successfully despite several attempts, these and even perhaps, AutoML (Machine learning using cloud computing), should be considered in future prediction of oil well failures.

2. **Implement XGBoost for Predictive Modelling** (High Priority):
    - Given its high accuracy and reliability, XGBoost can be the primary model for predicting oil well failures. This model can help in making informed decisions and reducing downtime.
3. **Utilize Decision Tree for Interpretability** (Medium Priority):
    - The Decision Tree model, while slightly less accurate than XGBoost, provides valuable insights into the decision-making process. This model is recommended for scenarios where understanding the underlying factors is crucial.

4. **Future Analysis on ESP Artificial Lift System and Run Life** (High Priority):
    - Future analysis on oil well failure should include a deeper investigation of the possibility of the ESP Artificial Lift System having longer run life. There should also be future analysis on injector wells and the relationship between the injectors and the production wells.
5. **Regular Model Validation and Update** (Medium Priority):
    - Continuous validation and updates of the predictive models are essential to maintain accuracy over time. Cross-validation techniques should be used to assess model performance periodically.

- **Implementation strategies**
    1. **Data Integration and Cleaning:**
        - Ensure consistent data quality by standardizing formats and addressing missing values through imputation techniques. This will facilitate better integration of new data and improve model reliability.
    2. **Deploying Predictive Models:**
        - Develop a deployment strategy that includes regular updates and monitoring of the models. Integrate the models into the company's existing maintenance management systems to provide real-time failure predictions.
    3. **Training and Support:**
        - Provide training for field technicians and decision-makers on how to interpret and use the model predictions. This will ensure that the insights are effectively utilized in operational decision-making**.**

- **Potential benefits**
    1. **Operational Efficiency:**
        - Improved prediction accuracy can lead to timely maintenance actions, reducing downtime and enhancing overall operational efficiency.

2. **Cost Savings:**
    - Proactive maintenance based on model predictions can significantly reduce the costs associated with unplanned interventions and unexpected equipment failures.
3. **Environmental Sustainability:**
    - By preventing potential failures, the models can help in avoiding incidents that could lead to environmental damage, thus supporting sustainable operations.

- **Monitoring and evaluation**
    1. **Performance Metrics**:
        - Use key performance indicators (KPIs) such as model accuracy, precision, recall, and F1 score to continuously monitor the effectiveness of the predictive models. Regularly use the Learning Curve to ensure there is no overfitting or underfitting of the model.
    2. **Feedback Loop**:
        - Establish a feedback loop with field technicians to gather insights on model performance and areas for improvement. This will help in refining the models and addressing any practical challenges encountered during deployment.
- **Continued analysis or research**

    1. **Exploring Alternative Models**:
        - Investigate other machine learning algorithms such as Random Forest and Support Vector Machine to compare performance and potentially enhance predictive accuracy.
    2. **Longitudinal Studies**:
        - Conduct longitudinal studies to understand the long-term effectiveness of the models and their impact on operational performance and cost savings.
    3. **Industry Collaboration**:

- Collaborate with other companies in the oil and gas industry to share insights and best practices, fostering innovation and continuous improvement in predictive maintenance techniques.

By implementing these recommendations and strategies, the organization can maximize the value derived from the predictive models, leading to significant improvements in operational efficiency, cost savings, and environmental sustainability.

Ethical Disclaimer

This project was conducted with a strong commitment to adhering to key ethical principles, ensuring transparency, and maintaining the highest standards of integrity throughout its duration. Our ethical considerations encompassed data privacy, informed consent, fairness, and non-discrimination.

Adherence to Key Ethical Principles

1. Data Privacy:

- All data used in this project were anonymized to protect the privacy of the company and entities involved. Sensitive information, such as well identifiers and geographic locations, was removed to prevent any potential breach of confidentiality.

2. Informed Consent:

- Data collection methods were reviewed to ensure that all participants or data sources were informed about the purpose of the project and provided consent for their data to be used in this research. (Negm, 2023 - 2024)

3. Fairness and Non-Discrimination:

- The project team was committed to ensuring that the predictive models developed were free from biases and did not discriminate against any group. Efforts were made to use balanced datasets and apply techniques to mitigate data imbalance, such as SMOTE (Synthetic Minority Over-sampling Technique).

Specific Actions Taken

1. Data Anonymization:

- Implemented robust data anonymization techniques to safeguard personal and sensitive information, ensuring compliance with data protection regulations and ethical standards. (Negm, Data Anonymization Technique, 2023 - 2024)

2. Transparency in Methodology:

- Maintained transparency in all research methodologies, including the CRISP-DM framework, to ensure that the processes and decisions were clear and accountable. Detailed documentation of data preparation, modeling, and validation steps was maintained to provide a transparent audit trail.

3. Bias Mitigation:

- Conducted regular reviews and analyses to identify and address any potential biases in the data and models. This included using diverse datasets and applying statistical techniques to ensure fair and unbiased outcomes. (Negm, 2023 - 2024)

Acknowledgment of Limitations

Despite our best efforts to uphold these ethical principles, we acknowledge that limitations might still exist. These limitations could include potential biases inherent in the original data, unforeseen impacts of data anonymization on the analysis, and the challenges of ensuring complete fairness and non-discrimination in complex predictive models.

Commitment to Continuous Improvement

We are committed to continuous improvement in our ethical practices. We will regularly review and update our methodologies to align with evolving ethical standards and best practices. Feedback from stakeholders and advancements in ethical guidelines will be incorporated to enhance the integrity and fairness of our research.

By adhering to these ethical principles and taking specific actions to uphold them, we aim to ensure that our research is conducted responsibly and with the utmost respect for the rights and privacy of all individuals and entities involved.

References

Jahir Guiterrez, T. O. (2024, June 11). Modelling Analysis_XGBoost. Saint John, New Brunswick, Canada.

Jahir Gutierrez, T. O. (2024, June 06). Modeling Analysis_Decision Tree. Saint John, New Brunswick, Canada.

Jahir Gutierrez, T. O. (2024). *Modelling Analysis Report - Predictive Analytics for Oil Well Failures: A Machine Learning Approach*. Saint John.

Negm, H. (2023 - 2024, January - April). A data ethics checklist - Data Collection. Saint John, New Brunswick, Canada.

Negm, H. (2023 - 2024, January - April). Data Anonymization Technique. Saint John, New Brunswick, Canada.

Titilola Oduwole, J. G. (2024, June 06). Modeling Analysis_LogisticRegression. Saint John, New Brunswick, Canada.

Titilola Oduwole, J. G. (2024). *Project Proposal - Predictive Analytics for Oil Well Failures: A Machine Learning Approach*. Saint John.

Titilola Oduwole, J. G. (2024, June). Statistical Analysis and Data Visualization. Saint John, New Brunswick, Canada.