Rockwell Automation

Data Science and Analytics Projects Portfolio

Domain Expert Team

Report prepared by
Abhinav Sharma
3 January 2024

Rockwell Automation India Pvt. Ltd.

Tel.: +91-120-467 1111
Fax: +91-120-421 7932
raindia@ra.rockwell.com

Date: 03-Jan-24

To Whomsoever It May Concern

This is to certify that Mr. Abhinav Sharma a student at IIT Guwahati undertook a project on “price prediction model, Anamoly detection and Soft sensor models based on AI/ML” at Rockwell Automation India Pvt Ltd from 01-Dec-23 to 31-Dec-23.

Mr. Abhinav has successfully completed the project under the guidance of Mr. Amit Gupta, Consultant Manager. He is a sincere and hard-working student with desire to learn new things. His performance and conduct during the tenure were found satisfactory.

We wish him success in all future endeavors.

For **Rockwell Automation India Private Limited.**

Pallav Purkayastha
(Manager- Talent Acquisition India)

Project 1: Price Prediction of Sale of Product X

Overview:

The objective of this project was to predict the sale price of Product X. The initial dataset was provided in xlsv format, which was converted to csv for ease of manipulation. The project involved data cleaning and observation, followed by an attempt to apply a neural network (NN) on 12 features, resulting in an unsatisfactory accuracy of 14%.

Iterative Approach:

1. **Feature Selection:**
   - Initially attempted to use 12 features, resulting in poor accuracy.
   - Guidance from the team lead led to focusing only on date and target value (DS-y format).
2. **Time Series Analysis:**
   - Divided the previous 3-year data into 2 years and 1 year segments.
   - Applied Fbprophet, ARIMA, SARIMA, and LSTM NN.
   - Evaluated models using RSME and KL divergence.
   - Selected Fbprophet due to superior performance and used it to predict sales for the next 2 years.

Outcome:

The final model, based on Fbprophet, demonstrated significantly improved accuracy compared to the initial NN attempt. This approach is expected to enhance sales prediction accuracy for the next 2 years.

Root Mean Squared Error (RSME):

RSME is a commonly used metric to measure the average magnitude of the errors between predicted and actual values. It is calculated as the square root of the mean of the squared differences between the predicted (P_i) and actual (A_i) values.

$$RSME = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - A_i)^2}$$

- n: Number of observations in the dataset.
- $\sum$: Summation over all observations.

The lower the RSME value, the better the model's predictions align with the actual values.

Kullback-Leibler (KL) Divergence:

KL Divergence is a measure of how one probability distribution diverges from a second, expected probability distribution. In the context of time series analysis, it is often used to compare the predicted and actual probability distributions.

$$KL(P||Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right)$$

- $P(i)$: Probability of the i-th observation in the actual distribution.
- $Q(i)$: Probability of the i-th observation in the predicted distribution.

The KL Divergence measures the information lost when the predicted distribution is used to approximate the actual distribution. A lower KL Divergence indicates a better match between the two distributions.

The KL Divergence measures the information lost when the predicted distribution is used to approximate the actual distribution. A lower KL Divergence indicates a better match between the two distributions.

Interpretation:

- **RSME:** A lower RSME suggests that the model's predictions are closer to the actual values, indicating better accuracy.
- **KL Divergence:** A lower KL Divergence indicates a closer match between the predicted and actual distributions, signifying a more accurate model.

In the context of the Price Prediction project, the evaluation based on RSME and KL Divergence helped in selecting the Fbprophet model over other algorithms. The chosen model demonstrated superior performance in capturing the underlying patterns in the sales data, contributing to better accuracy for future predictions.

Conclusion and Impact:

1. **Project Impact:**
   - The project concludes with the successful development and deployment of an accurate model for predicting the sale price of Product X.
2. **Practical Application:**
   - The selected Fbprophet model is deemed suitable for practical use in an industrial setting, where accurate sales predictions are crucial for strategic planning.
3. **Resource Optimization:**
   - The accurate prediction capabilities of the model can contribute to resource optimization and financial planning, addressing challenges faced in the industrial domain.

Project 2: Soft Sensing Model

Overview:

The Soft Sensing Model project aimed to develop an efficient model for industrial applications to predict certain parameters. The initial dataset, provided in CSV format, underwent preprocessing steps, including the handling of multiple NaN values, normalization using min-max scaling, and feature reduction.

Data Preprocessing:

1. **NaN Value Handling:**
   - Implemented strategies for multiple NaN value deletion.
2. **Normalization:**
   - Applied min-max scaling to normalize the dataset.
3. **Feature Reduction:**
   - Reduced the initial set of 15 features to a more manageable 5 features.
   - Applied Mutual Information (MI) on the training dataset to select the top 5 features with the highest MI.
4. **Dimensionality Reduction (PCA):**
   - Utilized Principal Component Analysis (PCA) to further reduce dimensionality.

The formula essentially measures how much knowing the value of one variable reduces the uncertainty about the other. In the context of feature selection, the mutual information between a feature (Xi) and the target variable (y) can be expressed as:

$$MI(X_i, y) = \sum_{x_i \in X_i} \sum_{y \in y} p(x_i, y) \log \left(\frac{p(x_i, y)}{p(x_i) \cdot p(y)} \right)$$

This MI score is calculated for each feature (Xi) in the dataset concerning the target variable (y). Higher MI scores indicate a stronger association between the feature and the target variable.

Model Development:

1. **Train-Test Split:**
   - Split the dataset into a training set (80%) and a testing set (20%) with a random state of 42.
2. **Model Selection:**
   - Applied multiple models, including KNN and Random Forest.
   - Achieved a 68% accuracy with KNN and 64% with Random Forest.
3. **Model Refinement:**
   - Explored feature adjustments and PCA modifications to enhance Random Forest accuracy to 70%+.
4. **Weight and Bias Calculation:**
   - Determined the weight and bias of the optimized Random Forest model.

Feature Engineering and Analysis:

1. **Feature Dependency Analysis:**
   - Explored the dependency of the predicted variable on the top 5 features with the highest mutual information.
   - Applied the formula $y = \sum_{i=1}^{5}(mi \cdot xi) + c$ to analyze feature dependencies.
2. **Variable Ranking:**
   - Assigned variable rankings based on their importance in predicting the target variable.
3. **Correlation Analysis:**
   - Evaluated the correlation between predicted and actual values.

Industry Application:

1. **Accuracy in Industrial Context:**
   - Achieved a very good accuracy range of 60-70%, suitable for industrial applications.
2. **Optimization for Practical Use:**
   - Explored changes in top features and PCA to enhance Random Forest accuracy to 70%+ for industry-specific requirements.
3. **Feature Engineering for Predictive Power:**
   - Applied feature engineering techniques to boost the predictive power of the model.
4. **Model Deployment:**
   - Selected the best-performing model with the least RSME for deployment in an industrial setting.

Conclusion:

The Soft Sensing Model project successfully addressed challenges in preprocessing, feature selection, and model optimization. The resulting model demonstrated high accuracy in an industrial context, making it a valuable tool for predictive analytics in practical applications. The comprehensive feature analysis and engineering contribute to its effectiveness in real-world scenarios.

Project 3: Anomaly Detection

Overview:

Project 3 focused on the detection of anomalies in machine-generated data. The dataset was initially in xlsv format, which was converted to csv for ease of manipulation. Missing data was removed from the dataset as a sufficient amount of data was available for analysis. Anomaly detection, a crucial aspect in industrial applications, was approached as a classification problem to determine whether the machine exhibited anomalies.

Data Preprocessing:

1. **File Format Conversion:**
   - Converted the dataset from xlsv format to csv for ease of handling.
2. **Missing Data Removal:**
   - Eliminated missing data from the dataset, leveraging the availability of a substantial amount of data.

Model Deployment:

1. **Artificial Neural Network (ANN):**
   - Deployed an Artificial Neural Network (ANN) as the primary model for anomaly detection.
   - Configured the ANN architecture for classification:
     - Input layer with 64 neurons.
     - Hidden layers with 32 and 16 neurons, respectively.
     - Output layer for classification.
     - Implemented dropout with a probability of 0.5 to enhance generalization.
2. **Random Forest:**
   - Utilized Random Forest as the secondary model for anomaly detection.
   - Configured the Random Forest architecture as the second-highest performing model.

Problem Statement:

- **Classification Problem:**
  - Framed the anomaly detection task as a classification problem.
  - The models aimed to determine whether the machine exhibited anomalies based on the input data.

Industrial Application:

- **Deployment at Industrial Level:**
  - Deployed the anomaly detection models at an industrial level.
  - The classification model proved valuable in distinguishing between normal and anomalous machine behavior.
- **Financial Resource Saving:**
  - Addressed the practical significance of anomaly detection by highlighting its role in saving financial resources.
  - Detection of anomalies can lead to timely interventions, minimizing potential financial losses in an industrial setting.

Conclusion:

Project 3 successfully implemented anomaly detection using Artificial Neural Networks (ANN) and Random Forest models. By framing the anomaly detection task as a classification problem, the models demonstrated effectiveness in distinguishing normal and anomalous machine behavior. The industrial-level deployment highlighted the practical implications of the models in saving financial resources through early detection and intervention.