Ada Tech

Programa Data4All

Business Case

Relatório Modelo Preditivo

Consumo mensal de energia em fábrica

Amanda Borges Matos Santana Magalhães

Cuiabá - MT

2023

Ada Tech

Programa Data4All

Business Case

Relatório Modelo Preditivo

Consumo mensal de energia em fábrica

Relatório do Projeto Preditivo do Consumo Mensal de Energia em uma Fábrica desenvolvido para o Programa Data4All da Ada Tech, como um Business Case.

Amanda Borges Matos Santana Magalhães

Cuiabá - MT

2023

Conteúdo

1 Resumo

Este relatório aborda o modelo preditivo do consumo mensal de energia. A priori foi construido o Machine Learning Canvas, os 5w e o problema de negocio. A partir disso, foi contruída a análise exploratória dos dados e foram identificadas as melhores features para serem utilizadas no modelo. Ainda foi abordado sobre as outliers encontradas nos dados, sejam por erro de entrada de informação ou falta de algum componetente explicativo da situação.

Após a análise exploratória de dados, foi realizado o modelo preditivo, sendo estes três modelos de regressão: Regressão KNN (K-Nearest Neighbors), Regressão Linear e Random Forest.

No relatório, também são discutidas métricas de avaliação de desempenho, como o erro quadrático médio (MSE) e o coeficiente de determinação (R^2), que podem ser usados para avaliar o desempenho dos modelos e sua capacidade de fazer previsões precisas.

Em resumo, este relatório fornece uma visão geral do projeto. Explora os princípios fundamentais e o funcionamento dos modelos, destacando suas aplicações, vantagens e limitações.

2 Introdução

Este relatório apresenta o projeto da predição do consumo de energia em uma fábrica. Quanto a análise exploratória de dados, objetivo desta é compreender a eficiência energética da fábrica, especialmente em relação ao fator de potência.

Corrente principal e corrente atrasada: No circuito elétrico, a corrente principal é aquela utilizada para alimentar os equipamentos elétricos e realizar trabalho útil. Já a corrente atrasada é a corrente que flui em resposta aos elementos capacitivos e indutivos presentes no circuito elétrico.

Fator de potência: É uma medida da eficiência com que a energia elétrica é utilizada. Um fator de potência baixo indica a presença de uma grande componente de corrente atrasada no circuito elétrico, o que resulta em desperdício de energia elétrica e custos mais elevados na conta de energia. Por outro lado, um fator de potência alto indica um uso mais eficiente da energia elétrica, com menos desperdício.

Carga elétrica: É a quantidade de energia elétrica consumida por um dispositivo ou circuito elétrico. Em um circuito com cargas capacitivas ou indutivas, a corrente atrasada é uma componente significativa da corrente total, resultando em um baixo fator de potência.

Medição de CO^2: Através da medição do CO2 associado ao consumo de energia, é possível identificar ineficiências energéticas. Ao medir o CO2 relacionado ao consumo de energia, é possível identificar áreas de baixa eficiência energética e altas emissões de CO2. Com base nesses dados, podem ser implementadas medidas de eficiência energética, como a atualização de equipamentos obsoletos, o uso de tecnologias mais eficientes e a adoção de práticas de conservação de energia. Isso pode levar a uma redução significativa nas emissões de CO2.

A Random Forest é um modelo de aprendizado de máquina baseado em árvores de decisão. É uma abordagem de ensemble, onde várias árvores de decisão são construídas e combinadas para fazer previsões mais precisas. Cada árvore na floresta é treinada em uma amostra aleatória do conjunto de dados e, durante a previsão, as respostas das árvores individuais são combinadas para obter uma previsão final.

A Regressão Linear é um modelo paramétrico que estabelece uma relação linear entre uma variável dependente e uma ou mais variáveis independentes. O modelo busca encontrar os coeficientes de regressão que minimizam a soma dos erros quadrados entre os valores observados e os valores preditos.

A Regressão KNN é um algoritmo de aprendizado de máquina não paramétrico que utiliza a ideia de que pontos de dados semelhantes tendem a ter valores semelhantes. O algoritmo encontra os k vizinhos mais próximos

de um novo ponto de dados com base na distância euclidiana e usa os valores desses vizinhos para prever o valor do novo ponto. A escolha adequada do valor de k é importante para o desempenho do modelo, e a normalização das características dos pontos de dados pode ser necessária para evitar viés.

3 Metodologia

A seguir, descreveremos os principais passos e etapas do desenvolvimento do modelo preditivo:

3.1 Pré-processamento dos dados

- Importação das bibliotecas necessárias, incluindo pandas, numpy, matplotlib e outras.
- Carregamento o conjunto de dados contendo informações sobre consumo de energia elétrica, corrente, fator de potência e quantidade de CO2.
- Realização de análise exploratória dos dados para entender a distribuição e identificar outliers.
- Para realizar a análise, foram utilizados dados históricos do consumo de energia elétrica na fábrica, incluindo informações como data, hora, consumo em kilowatt-hora (kWh), fator de potência e quantidade de CO2 emitida. Os dados foram coletados ao longo de um período de um ano e abrangem todas as áreas da fábrica.
- Identificação de outliers nos dados de consumo de energia e removendo esses valores, exceto os do mês de dezembro, que será usado para teste do modelo.

3.2 Separação dos dados de treino e teste

- Divisão dos dados em conjuntos de treino e teste.
- Selecão das features relevantes para o modelo, incluindo corrente atrasada, corrente principal, fator de potência atrasado, fator de potência principal e quantidade de CO2.

3.2.1 KNN Regression

- Aplicação da técnica de padronização dos dados utilizando o StandardScaler.
- Realização da escolha do valor adequado de k vizinhos para o modelo, utilizando a curva de aprendizado.

- Treinamento do modelo KNN Regression com os dados de treino padronizados.
- Avaliação do desempenho do modelo utilizando as métricas MSE (Mean Squared Error), R^2 (R-squared) e R^2 ajustado.

3.2.2 Linear Regression

- Padronização os dados de treino e teste utilizando o StandardScaler.
- Treinamento do modelo de regressão linear com os dados de treino padronizados.
- Avaliação do desempenho do modelo utilizando as métricas MSE, R^2 e R^2 ajustado.

3.2.3 Random Forest Regression

- Realização da escolha do número de estimadores utilizando a curva de aprendizado e valores de R^2.
- Treinamento do modelo de random forest regression com os dados de treino.
- Avaliação do desempenho do modelo utilizando as métricas MSE, R^2 e R^2 ajustado.

3.3 Resultados e Avaliação

Após a execução do modelo preditivo utilizando KNN Regression e Linear Regression, obtivemos os seguintes resultados:

3.3.1 KNN Regression

- MSE (Mean Squared Error) - Treinamento: 1.1828699103410798
- MSE - Teste: 1.1022945926789318
- R^2 - Treinamento: 0.9989408149679188
- R^2 - Teste: 0.998874158723009
- R^2 ajustado - Treinamento: 0.9989406378407519
- R^2 ajustado - Teste: 0.9988722633673238

3.3.2 Linear Regression

- MSE - Treinamento: 21.635383725067417
- MSE - Teste: 12.749830126279267
- R^2 - Treinamento: 0.980626885167517
- R^2 - Teste: 0.9869778141649932
- R^2 ajustado - Treinamento: 0.980623645407854
- R^2 ajustado - Teste: 0.986955891293217

3.3.3 Random Forest Regression

- MSE - Treinamento: 0.1079634025595289
- MSE - Teste: 0.6131030342142538
- R^2 - Treinamento: 0.9999033256159414
- R^2 - Teste: 0.9993738001551026
- R^2 ajustado - Treinamento: 0.9999033094491158
- R^2 ajustado - Teste: 0.9993727459466095

4 Sobre os modelos

4.1 Linear Regression

O modelo de Regressão Linear é uma técnica estatística utilizada para modelar a relação entre uma variável dependente (alvo) e uma ou mais variáveis independentes (features). Ele se baseia na suposição de que existe uma relação linear entre as features e o alvo, ou seja, é esperado que as mudanças nas features estejam linearmente relacionadas às mudanças no alvo.

A representação matemática básica de um modelo de Regressão Linear é dada por:

$$y + B1x1 + B2x2 + ... + Bnxn + E$$

onde:

- y é a variável dependente (alvo);
- x^1, x^2, ..., x são as variáveis independentes (features);
- B°, B^1, B^2, ..., Bn são os coeficientes de regressão que representam o efeito de cada feature no alvo;
- E é o termo de erro, que representa a parte da variabilidade do alvo que não é explicada pelas features.

O objetivo do modelo de Regressão Linear é estimar os valores ótimos dos coeficientes de regressão (, , , ...,) de modo a minimizar a soma dos quadrados dos resíduos (ou erros), que é a diferença entre os valores reais do alvo e os valores previstos pelo modelo.

A estimativa dos coeficientes de regressão é geralmente feita por meio do método dos mínimos quadrados ordinários (MMQ), que encontra a linha reta que melhor se ajusta aos dados ao minimizar a soma dos quadrados dos resíduos. O MMQ calcula os coeficientes que minimizam a seguinte equação:

$$Somatório(y - \hat{y})^2$$

onde:

- y são os valores reais do alvo;
- $\hat{y}$ são os valores previstos pelo modelo para cada observação i.

4.2 Random Forest Regression

A Random Forest Regression (Regressão com Floresta Aleatória) é um modelo de aprendizado de máquina que combina várias árvores de decisão para realizar previsões. A parte matemática e estatística desse modelo envolve os seguintes conceitos:

Árvores de decisão: Uma árvore de decisão é uma estrutura hierárquica composta por nós (ou pontos de divisão) e folhas (ou nós terminais). Cada nó interno representa uma decisão com base em uma feature específica e um ponto de divisão, que divide o espaço de dados em subconjuntos mais homogêneos. Cada folha representa uma previsão do valor do alvo.

Floresta aleatória: A Random Forest Regression é uma coleção (ou ensemble) de várias árvores de decisão independentes. Cada árvore é treinada em um subconjunto aleatório dos dados de treinamento e usa um subconjunto aleatório das features. A combinação das previsões de todas as árvores é usada para obter uma previsão final. Estatisticamente, essa abordagem ajuda a reduzir a variância e o overfitting, fornecendo resultados mais estáveis e precisos.

Durante o processo de treinamento da Random Forest Regression, cada árvore de decisão é treinada usando a técnica de bagging. O bagging envolve a amostragem com reposição dos dados de treinamento para criar várias amostras de treinamento diferentes. Em seguida, cada árvore é treinada em uma amostra diferente. Essa abordagem aumenta a diversidade das árvores e promove um melhor desempenho geral do modelo.

A Random Forest Regression fornece uma medida estatística útil para avaliar a importância das features. Essa medida é baseada na diminuição média da impureza (ou ganho de informação) que cada feature fornece ao modelo. Quanto maior for a diminuição da impureza, maior será a importância da feature para as previsões do modelo.

4.3 KNN Regression

O modelo KNN Regression (K-Nearest Neighbors Regression) é um algoritmo de aprendizado de máquina usado para realizar previsões com base na proximidade dos vizinhos mais próximos em um conjunto de dados.

O modelo KNN Regression utiliza a ideia de vizinhança para realizar previsões. Para cada ponto de dados no conjunto de treinamento, o modelo identifica os K pontos mais próximos com base em uma medida de distância, geralmente a distância euclidiana. A escolha do valor K é um hiperparâmetro que precisa ser definido antes do treinamento.

Em KNN Regression, é comum ponderar os valores dos vizinhos próximos

com base em sua proximidade. Ou seja, os vizinhos mais próximos têm um peso maior em relação aos vizinhos mais distantes. Diferentes abordagens podem ser usadas para atribuir pesos, como o inverso da distância ou o inverso do quadrado da distância.

Para fazer uma previsão, o modelo KNN Regression calcula a média ponderada dos valores dos vizinhos mais próximos. Os pesos atribuídos a cada vizinho são usados para determinar a contribuição deles para a previsão final. Essa média ponderada é o valor previsto para o ponto de dados de interesse.

A distância euclidiana é uma medida comumente utilizada para calcular a distância entre dois pontos em um espaço multidimensional. Ela é calculada como a raiz quadrada da soma dos quadrados das diferenças entre as coordenadas dos dois pontos. A distância euclidiana é usada no modelo KNN Regression para medir a proximidade entre pontos de dados e identificar os vizinhos mais próximos.

Além disso, esses modelos possuem medidas estatísticas para avaliar a qualidade do ajuste e a significância estatística dos coeficientes. A

O MSE é uma medida estatística comumente utilizada para avaliar a qualidade do ajuste da Random Forest Regression. Ele calcula a média dos quadrados dos erros entre as previsões do modelo e os valores reais do alvo. Quanto menor for o MSE, melhor será o ajuste do modelo aos dados de treinamento.

Coeficiente de determinação R^2, é outra medida estatística usada para avaliar o desempenho dos modelos. Ele indica a proporção da variabilidade do alvo que é explicada pelas features do modelo. O R^2 varia de 0 a 1, sendo 1 indicativo de um ajuste perfeito.

E ainda possui o R^2 ajustado, que pune do coeficiente de acordo com as quantidades de features utilizadas no modelo.

5 Conclusão

Com base nos resultados obtidos, podemos concluir que tanto o modelo KNN Regression, Random Forest Regression, quanto o modelo Linear Regression apresentaram um desempenho satisfatório na tarefa de previsão do consumo de energia elétrica. Ambos os modelos foram capazes de capturar as relações entre as features de entrada e a variável alvo, resultando em baixos erros de previsão e bons coeficientes de determinação (R^2) tanto nos conjuntos de treinamento quanto nos conjuntos de teste.

No entanto, cada modelo possui suas características e limitações específicas. Portanto, a escolha do modelo final dependerá de outros critérios, como interpretabilidade, simplicidade e requisitos específicos da empresa, por este motivo, foram mantidos os três modelos no projeto final.

Ainda é necessárias análises adicionais e validações para aprimorar ainda mais o desempenho do modelo preditivo.