Relatório # 1

Análise e preparação de dados para modelagem de previsão de geração em usinas fotovoltaicas

Resumo

- Este relatório detalha as etapas de entendimento e preparação dos dados para desenvolvimento de um modelo de previsão de geração dirigido por dados para usinas fotovoltaicas.
- Inicialmente, foram examinadas as fontes de dados, com destaque para as medições coletadas da estação solarimétrica e dos inversores de frequência.
- Diversas atividades foram realizadas na etapa de entendimento dos dados, desde a coleta inicial até a exploração e verificação dos dados.
- Além disso, foram identificados problemas de qualidade dos dados, como intervalos sem dados e *outliers*, os quais foram tratados durante a preparação dos dados.
- As etapas de seleção, limpeza, construção, integração e formatação dos dados foram desenvolvidas, culminando na criação de um *script* capaz de gerar o conjunto final de dados que alimentará o modelo de previsão.

1. Entendimento dos dados

- O objetivo desta etapa foi examinar as fontes de dados para o projeto.
- Foi realizada uma coleta inicial de dados, seguida por atividades para: familiarização com os dados; identificação de problemas de qualidade de dados; descoberta dos primeiros *insights* sobre os dados; e detecção de subconjuntos interessantes a fim de formar hipóteses para informações ocultas.
- Particularmente, o entendimento dos dados envolveu as seguintes atividades (que serão detalhadas nas próximas seções):
    1. Coleta de dados: localização e extração dos dados mais relevantes para o problema.
    2. Descrição dos dados: identificação do formato, quantidade e outras características dos dados.
    3. Exploração dos dados: navegação pelos dados em termos de pesquisa, visualização e reporte.
    4. Verificação dos dados: avaliação da qualidade dos dados.

1.1. Coleta de dados

- Os dados históricos de geração solar da UFV de Coromandel foram eleitos para a modelagem da previsão de geração.
- A UFV de Coromandel é composta por 2 unidades geradoras, e conta com 8 inversores em cada unidade (16 inversores ao todo). Além disso, há uma única estação solarimétrica na usina.
- Os dados particularmente relevantes são as medições coletadas da estação solarimétrica e as medições coletadas a nível dos inversores.
- Foi feita uma cópia do banco de regressão (PostgreSQL), e as medições históricas dos inversores e da estação solarimétrica foram extraídas com o auxílio das linguagens SQL e Python.

1.2. Descrição dos dados

- Os dados são números reais (*floats*) que representam as medições das variáveis (temperatura, irradiância, potência, etc), e são acompanhados por suas respectivas estampas de tempo (ano-mês-dia hora:minuto:segundo).
- A frequência de coleta dos dados é distinta, mesmo que para um mesmo equipamento. Por exemplo: no caso dos inversores, o período de amostragem da Geração Diária é de 5 minutos, enquanto o período de amostragem da Potência DC é de 30 minutos. Analogamente, no caso da estação solarimétrica, o período de amostragem das variáveis de Irradiância e Temperatura do Módulo é de 1 minuto, enquanto o período de amostragem das outras variáveis é maior (e distinto).
- As medições coletadas encontram-se particionadas em diversas tabelas, identificadas pela usina, ano e mês.
- Há dados coletados *desde agosto de 2023*.

1.3. Exploração dos dados

- As análises de correlação linear entre as variáveis, feitas em alguns subconjuntos selecionados, revelaram que:
    - As potências AC e DC são fortemente correlacionadas, o que é esperado de inversores em bom estado de funcionamento.
    - Há uma forte correlação entre a hora do dia e a Geração Diária, o que também é esperado, já que como medida acumulativa, aumenta ao longo do dia (até estabilizar com o pôr do sol).

    - A temperatura do módulo solarimétrico é fortemente correlacionada com a temperatura ambiente e com a irradiância.
    - As potências AC e DC são fortemente correlacionadas com a Irradiância e com a temperatura do módulo solarimétrico.
    - Há uma forte correlação entre as Gerações Diárias de todos os inversores.
    - Há uma forte correlação entre a Geração Diária e a Irradiância.
- Das análises de correlação não linear, observou-se conexão entre a hora do dia e as potências AC e DC, com picos no intervalo de maior intensidade solar.

1.4. Verificação dos dados

- Há diversos intervalos em que não há dados, para diferentes variáveis. Por exemplo, é comum que não haja dados para os inversores no período das 00:00 às 08:00.
- Há também dados que não fazem sentido, por exemplo, valores de Geração Diária no período das 00:00 às 06:00 iguais aos do final do dia anterior. Outro exemplo são valores negativos (próximos de 0) de Irradiância (possivelmente alguma má calibração).
- Há períodos em que não há dados por algumas horas, ou até mesmo dias.
- Nos subconjuntos analisados, foram identificados alguns *outliers* utilizando-se o método IQR (Intervalo entre Quartis).
- Não foram identificados valores negativos de potência (AC e DC), geração diária e geração total nos subconjuntos analisados.

2. Preparação dos dados

- O objetivo desta etapa foi decidir quais conjuntos de dados serão utilizados.
- Foram realizadas atividades para construir o conjunto final de dados (que alimentará a modelagem), a partir dos dados brutos iniciais.
- As tarefas de preparação de dados provavelmente serão realizadas várias vezes ao longo do projeto, uma vez que lidar com dados envolve muita iteração.
- Particularmente, a preparação dos dados envolveu as seguintes atividades (que serão detalhadas nas próximas seções):
    1. Seleção de dados: definição de quais dados serão incluídos ou excluídos.

2. Limpeza de dados: realização de ações para resolver problemas de qualidade dos dados.
3. Construção de dados: criação de novos atributos e registros.
4. Integração de dados: realização de ações como mesclagem e agregação para trabalhar com os dados.
5. Formatação de dados: realização de correções na ordem e sintaxe dos dados.

2.1. Seleção de dados

- Como descrito na seção 1.1, os dados selecionados foram as medições coletadas da estação solarimétrica e as medições coletadas a nível dos inversores.
- Particularmente, as seguintes variáveis independentes foram inicialmente escolhidas para a modelagem:
    - *POA Irradiance* [W/m²]
    - *Module Temperature* [˚C]
    - *Air Temperature* [˚C]
    - *DC Power* [W]
    - *AC Active Power* [kW]
    - *Total Yield* [kWh]
- A variável dependente (variável alvo), a ser prevista, será a variável *Daily Yield* [kWh], a partir da qual podem ser calculados outros indicadores.
- Há outras variáveis particularmente interessantes (inclusive dos *combiners*), que por simplicidade serão inicialmente desconsideradas.
- Dada a natureza sazonal da geração solar, o ideal seria utilizar 1 ano ou mais de dados para a modelagem. Contudo, dado o número relativamente pequeno de variáveis, espera-se que aproximadamente um mês de dados seja suficiente para previsões de curto prazo (poucos dias).

2.2. Limpeza de dados

- *Outliers* filtrados com base no método IQR.
- Dados inconsistentes substituídos, por exemplo:
    - Dados não nulos de Geração Diária substituídos por zeros durante a madrugada.
    - Dados negativos de Irradiância substituídos por zeros.

- Criada série temporal única, com intervalos uniformes de 5 minutos para todas as medidas.
- Intervalos sem dados substituídos por valores imputados via *machine learning*, utilizando o algoritmo *MissForest,* baseado no método *Random Forest* (outros métodos como KNN, interpolação linear, médias móveis etc., podem ser utilizados).

2.3. Construção de dados

- Localmente, foi criada uma única tabela com as variáveis selecionadas como colunas, além da coluna com as estampas de tempo.
- Os novos registros da tabela são os dados tratados pela operação de limpeza dos dados.
- No banco original, pode-se construir uma tabela semelhante à tabela *expected_data* para a geração prevista pelo modelo.

2.4. Integração de dados

- Os dados (de potência e geração) dos 16 inversores foram agregados (pela soma) em uma única coluna para cada variável.

2.5. Formatação de dados

- A sintaxe dos dados foi alterada localmente, com nomes descritivos para as variáveis, a fim de facilitar o trabalho com os dados (o banco original particiona as medidas por tabelas a cada novo mês, além de identificar as variáveis pelo *high_code*).
- A ordem, a quantidade, a frequência e a completude dos registros foram alterados localmente, de modo a preparar os dados para a modelagem.

3. Conclusão

- A análise e preparação dos dados são etapas cruciais para o sucesso de qualquer projeto de IA.
- Foi demonstrada a importância de compreender os dados disponíveis, identificar problemas de qualidade e realizar os devidos tratamentos.

- Este relatório serve como um guia detalhado do processo realizado, fornecendo uma base para futuras análises e desenvolvimentos neste contexto específico.

Referências

[1] CRISP-DM (*CRoss Industry Standard Process for Data Mining)*: https://www.datascience-pm.com/crisp-dm-2/

[2] *MissForest*: https://pypi.org/project/MissForest/

[3] *Missing Value Treatment - Advanced*: https://ishanjainoffical.medium.com/missing-value-treatment-advanced-methods-f7fa05ec0f39

[4] *Interquartile Range*: https://en.wikipedia.org/wiki/Interquartile_range