**Previsão de Produção de Energia Elétrica em Usinas Eólicas**

**Aluno: André Luis Andrade Machado**

A Previsão de Produção de Energia Elétrica era um tema até então restrito aos conhecimentos necessários para uma correta gestão de Produção e Demanda de Energia dentro de uma Rede Elétrica Nacional, mas que vem ganhando cada vez mais espaço com o crescimento de popularidade do Mercado Livre de Energia, virando até mesmo tema de uma Competição, como é o caso deste trabalho.

Neste breve documento, pretende-se evidenciar todas as etapas realizadas até o momento ao longo do projeto, apenas para fins de melhor organização e clareza no acompanhamento do mesmo.

1. **Pré-Processamento**

Antes de aplicar os dados em qualquer modelo, foi necessário inicialmente realizar um Pré-Processamento dos dados disponibilizados pelos organizadores da Competição. O conjunto de dados fornecido consiste em Previsões feitas para Três Variáveis Meteorológicas: Velocidade do Vento (Latitudinal e Longitudinal), Temperatura e Cobertura de Nuvens (Em Percentual). Cada uma dessas previsões é disponibilizada por quatro NWP’s (Numerical Weather Prediction) diferentes e realizadas em alguns horários diferentes. Sendo mais específico, temos o seguinte esquema:

* NWP1: Velocidades do Vento(U e V), para 100m de altura, e Temperatura para os horários {00h,06h,12h,18h};
* NWP2: Velocidades do Vento(U e V), para 100m de altura, para os horários {00h,12h};
* NWP3: Velocidades do Vento(U e V), para 100m de altura, e Temperatura para os horários {00h,06h,12h,18h};
* NWP4: Velocidades do Vento(U e V), para 10m de altura, e Cobertura de Nuvens para os horários {00h,12h}.

Além disso, as previsões são disponibilizadas em três dias diferentes Dois Dias Antes, Um Dia Antes e no Mesmo Dia do Horário a ser previsto.

Um detalhe importante a ser destacado é que os horários e dias citados acima se referem ao tempo em que a previsão foi realizada e não ao horário em que a previsão gerada corresponde em si. Quanto a este último atributo, tirando algumas exceções, as previsões correspondem a valores para cada hora do dia.

Esse conjunto de informações gera um Conjunto de Dados com 105 colunas, o que pode ser visto como um valor relativamente grande.

Desse ponto, dois questionamentos foram levantados:

1. Existem Valores Nulos nos Dados? É razoável pensar em preenchê-los?
2. Diversas colunas se referem a mesma Variável nos Dados? Não seria interessante pensar em agregá-las em uma Coluna só de forma a reduzir a dimensionalidade dos dados?

Com relação ao primeiro tópico, a possibilidade de preencher os valores foi avaliada inicialmente, porém devido ao fato de algumas colunas possuírem uma quantidade de valores nulos muito grandes, a hipótese de utilizar esta técnica foi descartada, ao menos inicialmente. Isso foi um pouco facilitado pelo fato do XGBoost (Descrito mais a frente) conseguir trabalhar com valores nulos. Essa etapa pode ser retomada durante o projeto da Rede LSTM, visto que isso pode se tornar um problema.

Por fim, com relação ao último tópico, a similaridade entre as diferentes previsões relacionadas foi analisada e confirmada através de alguns gráficos, e por fim, as colunas foram agregadas via Mediana. A Mediana foi escolhida basicamente devido a sua robustez maior a Outliers do que a Média.

1. **Feature Engineering**

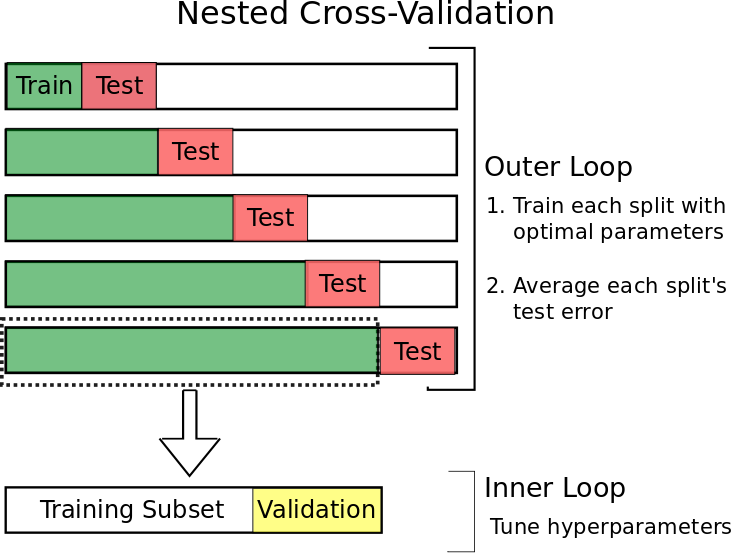
Nesta etapa, algumas Variáveis foram elaboradas a partir dos dados originais, visando aumentar a performance dos Modelos a serem criados. Abaixo, lista-se as Variáveis criadas:

* Magnitude e Direção do Vento (Em Graus);
* Valores a Uma Semana e Um Mês Atrás para as Variáveis Originais;
* Média e Variância por Mês Anterior das Variáveis Originais;
* Hora, Minuto, Dia da Semana e Dia do Ano aplicados a funções Sinusoidais;
* Distância (Em Horas) de cada Previsão para os Valores Máximos e Mínimos Globais
* Média e Variância Móvel em Janelas de 7 e 14 Amostras;
* Valor Máximo em Janela Expansível.

Por fim, para selecionar as Variáveis mais relevantes, o Método de Feature Selection conhecido como LOFO (Leave One Feature Out) foi implementado. Esse método basicamente consiste em avaliar o desempenho de um modelo retirando uma variável de cada vez e avaliando a sua influência no resultado ao ser removida. As variáveis que, ao serem removidas, influenciaram mais negativamente o desempenho do modelo são consideradas as mais importantes. Dentro dessa lógica, as variáveis foram avaliadas usando um modelo de XGBoost simples. Vários valores de limites de variáveis a serem selecionadas foram testadas, mas de fato, o melhor modelo obtido envolveu o uso de todas as variáveis, isto é, as variáveis originais junto às variáveis criadas.

1. **XGBoost**

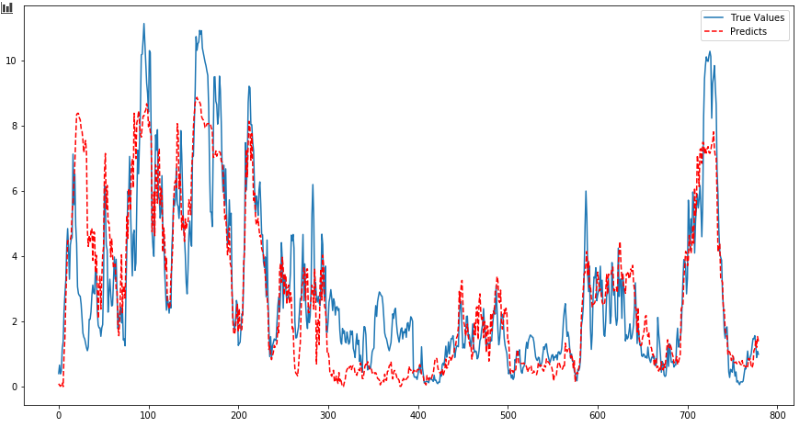
O primeiro modelo a ser considerado foi o XGBoost, que é uma implementação bastante utilizada de uma Gradient Boosted Decision Tree. Para o Ajuste dos Hiper parâmetros, foi utilizado um Método de Validação conhecido como Nested Cross Validation, onde os dados são separados em fold não aleatórios (Visto que os dados se tratam de Séries Temporais) e cada iteração envolve a utilização cumulativa dos folds, como pode ser visto abaixo.



Nesse processo, foi monitorado o Root Mean Squared Error (RMSE) no conjunto de Treino e o Cumulated Absolute Percentage Error (CAPE) no conjunto de validação. Esta última medida corresponde à métrica de erro considerada na competição, e o seu valor é utilizado para Early Stopping do Modelo.

O processo de Validação descrito acima é executado apenas com uma Fazenda Eólica (A Fazenda 3 é escolhida por possuir os maiores níveis de Produção de Energia Eólica), e é usado como função a ser otimizada pela biblioteca HyperOpt, a fim de facilitar a busca dos melhores parâmetros.

Por fim, o melhor modelo obtido neste processo ainda é avaliado em um conjunto de dados de Holdout, separado do conjunto utilizados no Cross Validation. Nesse caso, os dados são treinados sobre todos os dados utilizados no Cross Validation. Os resultados para o melhor modelo obtido apresentam um CAPE de 32.88 aproximadamente e podem ser visualizados no gráfico comparativo abaixo.



Após a seleção do modelo com melhor resultado no conjunto de Holdout, o modelo escolhido é replicado para as demais Fazendas Eólicas e por fim, as predições finais são geradas para serem enviadas à competição.

1. **Rede LSTM**

Após algumas alterações no escopo do projeto, decididas após algumas falhas, o próximo modelo a ser construído consiste no uso de Redes Neurais do tipo LSTM. O modelo ainda está em fase inicial de desenvolvimento, mas espera-se trabalhar em um cenário parecido com o que foi feito com o XGBoost. Dessa forma, o mesmo conjunto de dados e Processo de Validação provavelmente serão utilizados, apenas com as devidas alterações necessárias para o correto funcionamento do modelo, que até o momento consistem em escalonamento dos dados (Onde será utilizado o MinMaxScaler, escalonando os dados no intervalo [-1,1]) e preenchimento de valores nulos. Estes dois processos devem ser feitos com cuidado para evitar alterações de dados que possuem escalas diferentes, como a direção do vento, que está em graus.

1. **Falhas no Projeto**

Uma das falhas principais do projeto, senão a maior, foi a tentativa de aplicar como modelo a utilização de um modelo ARIMA-GARCH aplicado ao próprio resultado desejado, a Produção de Energia, com uma forma de realizar uma Autoregressão que pudesse trazer informações relevantes que talvez não estivessem sendo captadas pelo XGBoost. Após várias tentativas, nenhum resultado satisfatório foi obtido. Acredito que provavelmente o modelo não conseguiu captar as complexidades dos dados em questão.

Além disso, algumas tentativas de Variáveis ou alterações na etapa do Processamento também foram tentadas, porém não geraram resultados melhores no Modelo XGBoost.