INF-0615 Aprendizado de Máquina Supervisionado I

Relatório 01

Perda da Qualidade de Ar pela Concentração de Monóxido de Carbono

Integrantes do grupo:
Daniel Noriaki Kurosawa
Eric Uyemura Suda
Fernando Shigeru Wakabayashi

Objetivo

Este trabalho se propõe a criar um modelo para a perda de qualidade do ar pela concentração de monóxido de carbono usando modelos de regressão, para o trabalho final da matéria INF-0615 Aprendizado de Máquinas Supervisionado do curso Mineração de Dados Complexos da Universidade Estadual de Campinas.

Introdução

O conjunto de dados do estudo é disponibilizado em um arquivo CSV e é composto **No**: Índice do exemplo na base de dados. É um atributo não preditivo e não deve ser utilizado durante o treinamento dos modelos. **Year**: Ano no qual a mensuração foi realizada.**Month**: Mês no qual a mensuração foi realizada. **Day**: Dia no qual a mensuração foi realizada. **Hour**: Hora em que a mensuração foi realizada. **PM2.5:** Concentração de PM2.5 no ar (em ug/m_3). **PM10**: Concentração de PM10 no ar (em ug/m_3). **SO2**: Concentração de SO_2 no ar (em ug/m_3). **NO2**: Concentração de N O_2 no ar (em ug/m_3). **O3**: Concentração de O_3 no ar (em ug/m_3). **TEMP:** Temperatura em graus Celsius no momento da coleta. **PRES**: Pressão Atmosférica (em hPa) no momento da coleta. **DEWP:** Ponto de Condensação da Água na região da coleta. **RAIN:** Precipitação da água na região da coleta (em mm).**WD**: Direção do vento no momento da coleta. **WSPM**: Velocidade do Vento. E também a variável que vamos predizer **Target**: Concentração de Monóxido de Carbono (CO) no ar em ug/m_3 .

Inicialmente avaliamos a volumetria dos nossos dados para os datasets de treino e teste fornecidos:

Tipo Data Set	Volumetria
Treino	244.582

Validação	61.147
Teste	76.434
Volumetria Total	382.163

Tabela 01 - Volumetria dos datasets fornecidos

Posteriormente verificamos a disjunção entre os conjuntos, assim garantindo que não existem registros duplicados entre os sets de treino e validação.

Para este dataset temos as variáveis que compõem a data de medição da informação e também uma variável categórica de direção do vento.

Utilizamos o One hot encoder para o vento e transformamos os atributos categóricos em vetores ortogonais em si, além disso retiramos o "N" para ajustar a os graus de liberdade.

De acordo com a função summary do R , o dataset não apresenta nenhuma feature sem anotações.

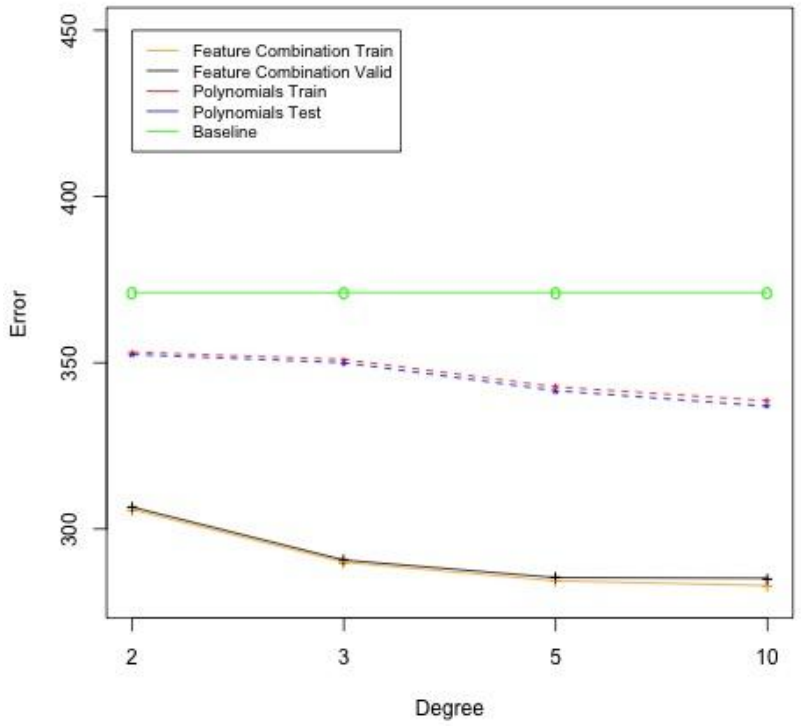
Ao analisar as features optamos por aplicar a normalização min-max , exceto para a variável "target" que será nossa variável a ser predita.

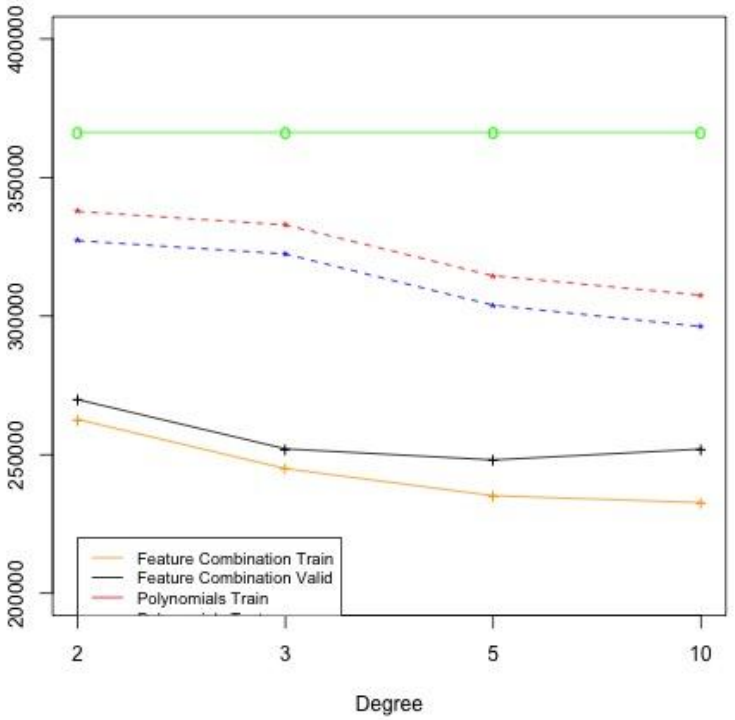
Depois da etapa normalização, criamos o modelo de base para comparar os resultados usando a regressão linear.

Para criação dos nossos modelos não consideramos a variável de chuva devido a alta concentração de valores em 0.

Criamos 4 modelos de combinação de features e 4 modelos de regressão ,com as potências de 2, 3, 5 e 10

Analisando os gráficos de MAE, MSE e R2, observamos que o erro de treino e de validação continuam caindo até o 4º grau, caracterizando underfitting. Assim, selecionamos o modelo de combinação à 5a potência pois este apresentou valores que consideramos suficientemente baixos de MAE , MSE e alto R2 no treino e na validação (ordens superiores de potência não adicionavam um ganho significativo de performance). Além disso, observamos overfitting no modelo ao observar que após a quinta potência o erro de treino teve uma diminuição, porém o da validação teve um aumento.





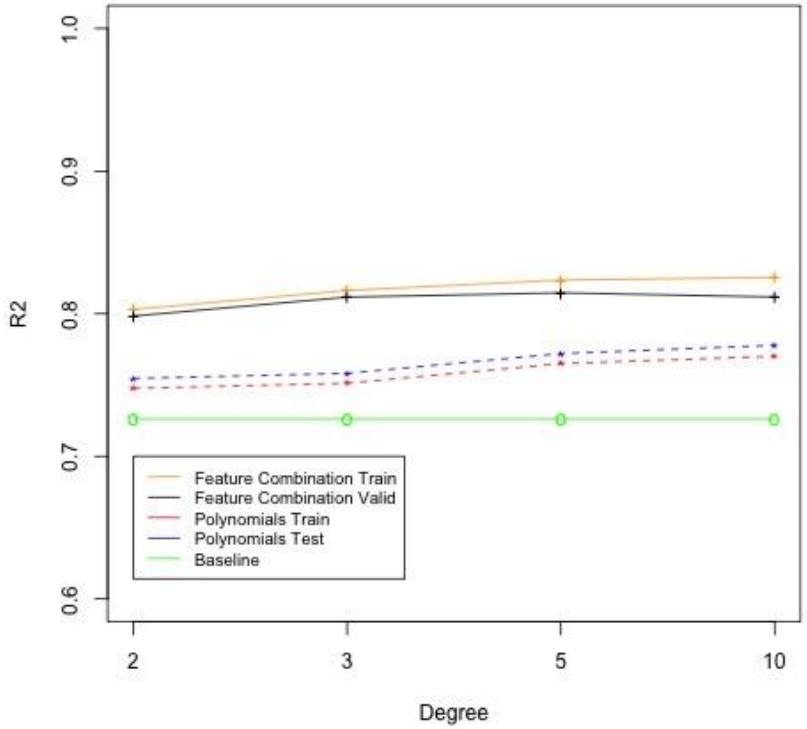


Gráfico 01- em sentido horário da esquerda para direita , MAE, MSE e R2

Resultados

1) Erro do baseline

Métrica	Valor
MAE	371.0654
MRE	366334.6
R2	0.7262772

2) Erro do melhor modelo de teste.

Usando o modelo combinado elevado à quinta potência, obtemos os seguintes valores para o conjunto teste.

Métrica	Valor
MAE	287,0232
MRE	24529,2
R2	0.8193355

Conclusão

Os 4 primeiros modelos utilizam combinação de variáveis levando em consideração a combinação entre as variáveis. Já os modelos polinomiais, não levam em consideração as interações entre os termos secundários. Devido a este fato, o primeiro grupo de modelos é mais sensível às interações entre as variáveis, o que resulta em um melhor desempenho dos modelos.