Machine Learning

ESTIMANDO VENDAS UNITÁRIAS DE PRODUTOS DE VAREJO

EQUIPE:

Alessandra Blucher

7° Semestre de Engenharia da Computação

Juliano Nehme Nassar

7° Semestre de Engenharia Mecatrônica

Sumário

Introdução 2

Objetivos 2

Metodologia 2

Cronograma 3

Referências 4

Introdução

Uma das ferramentas mais poderosas que podem ser utilizadas por empresas de varejo é a previsão de dados futuros, desde vendas, utilização e disponibilidade de recursos, movimentações de mercado e muito mais. Mesmo sendo algo essencial para definição de um planejamento, desde alocação de recursos até marketing guiado, muitas das empresas fazem previsões de maneira mal.

O desafio para deste projeto será criar uma ferramenta de previsão de vendas unitárias para produtos de varejo para lojas da empresa Walmart em 3 estados diferentes dos Estados Unidos. O intuito deste projeto é de participar da competição *M5 Forecasting - Accuracy*[1] e de compartilhar o conhecimento desenvolvido e estratégias utilizadas com todos que se interessarem.

Objetivos

Definir objetivos é uma tarefa difícil, mas necessária, os objetivos alinham bem as expectativas e ambições da equipe, além disso, servem para guiar o planejamento e cronograma para garantir uma entrega bem sucedida do projeto. Como a equipe é nova no assunto de previsão, ainda não é possível definir objetivos 100% concretos, mas já temos algumas ambições dados os estudos feitos até o momento, sendo elas:

- Aplicar técnicas já existentes de séries hierárquicas temporais
- Utilizar métodos clássicos de Aprendizado de Máquina e comparar seus resultados
- Conseguir acurácia melhor que métodos clássicos de predição (Que não envolvem aprendizado de máquina)
- Se bem sucedidos com estes métodos, tentar melhorar as previsões utilizando redes neurais (Ainda não foi definida uma topologia para esta etapa) (Extra 1)
- Ganhar a competição e pagar uma cerveja para o professor da matéria (Extra 2)

Metodologia

Quais métodos utilizar para este problema? O primeiro passo é definir se este problema é de regressão ou classificação, como é pedido a previsão da venda de produtos, e isto é uma quantidade numérica, o problema é de regressão.

O próximo passo é conseguir os dados e entender seu conteúdo, felizmente esta parte, que é uma das mais difíceis, já foi feita pelo Walmart e todos dados necessários são apresentados na página da competição, eles são divididos em 4 bases de dados:

- **calendar.csv:** Informações sobre as datas de venda (Feriado, eventos especiais, etc)
- **sales_train_validation.csv:** Dados do histórico de vendas de unidades de produtos por data e loja (Utilizado para treino e validação)
- **sell_prices.csv:** Informação sobre preço dos produtos por data e loja
- **sales_train_evaluation.csv:** Mesmas informações **sales_train_validation** mas utilizado para avaliação e resultado da competição

[1] https://www.kaggle.com/c/m5-forecasting-accuracy/overview/evaluation (Acessado dia 04/06/2020)

Além destas bases de dados ricas, é fornecido um padrão para a saída do modelo. Este modelo pode ser encontrado no arquivo **sample_submission.csv**, é importante se atentar a isso pois será fundamental entregar os dados na estrutura exigida pelo cliente.

O próximo passo será tratar estes dados, transformando as variáveis categóricas em Dummy variables e testando diferentes tipos técnicas para preparar dado, como separação estratificada, normalização, seleção de features se necessário, correlação, preenchimento de dados vazios com média, mediana ou moda, aplicar redução de dimensionalidade etc.

Depois de tratar os dados e ter toda está parte pronta, será necessário definir quais estratégias de HTS (Hierarchical time series) serão utilizadas, se possível, testar todas para enriquecer mais a discussão. As estratégias que a equipe deseja utilizar são Bottom-up, top-down, middle-out e optimal combination/reconciliation[2]. Definir estas estratégias é crucial pois é a partir disso que se como será feito o treinamento dos modelos de aprendizado de máquina (se será feito produto a produto, loja a loja, treinamento com todos dados, etc)

Com os dados tratados e primeira estratégia definida, serão iniciados os testes com diferentes modelos. Para um melhor entendimento dos dados, serão utilizados inicialmente modelos de regressão lineares, estes podem dar indícios de quais features são importantes, provavelmente não terão os melhores resultados, mas devem dar informações importantes.

Depois dos primeiros testes fazendo regressão e um pouco de regularização, pode-se seguir dois caminhos, tentar utilizar uma árvore de decisão para previsão de vendas (Mesmo sendo simples talvez funcione e é mais fácil de visualizar sua análise), ir direto para uma Random forest ou utilizar outras estratégias de modelos ensemble para procurar resultados melhores.

Após todos estes testes, se der tempo, serão feitas topologias de redes neurais, um possível teste seria uma rede recorrente ou tentar apresentar uma topologia diferente.

Para treinar os modelos e selecioná-los provavelmente será utilizada validação cruzada, como métrica inicial será utilizado o RMSE, se necessário serão inseridas nova métricas.

Cronograma

- ***09/06/2020*** – Leitura completa do livro ***Forecasting – principles and practices***, definir estratégia de HTS a ser utilizada, ter a pipeline para os dados pronta e ter iniciado os testes em regressões lineares.
- ***11/06/2020*** – Ter feito todos testes nos modelos de regressão linear e começar a retirar as conclusões deles. Já ter iniciado confecção do relatório.
- ***16/06/2020*** – Ter feito todos testes nos modelos de árvores de decisão e random forests e começar a retirar as conclusões deles. Definir uma estratégia para fazer um modelo Ensemble. Já ter iniciado confecção da apresentação
- ***18/06/2020*** – Ter feito todos testes no modelo Ensemble e começar a retirar as conclusões deles. Finalização do grosso da apresentação.
- ***Semana de provas*** – Mandar bem na apresentação (Com certeza né) e finalizar o relatório.

[2] https://medium.com/opex-analytics/hierarchical-time-series-101-734a3da15426 (Acessado dia 04/06/2020)

- ***25/06/2020*** – Entrega do relatório final.

Referências

https://www.kaggle.com/c/m5-forecasting-accuracy/overview (Acessado dia 04/06/2020)

https://medium.com/opex-analytics/hierarchical-time-series-101-734a3da15426 (Acessado dia 04/06/2020)

https://robjhyndman.com/papers/Hierarchical6.pdf (Acessado dia 04/06/2020)

https://otexts.com/fpp2/ (Acessado dia 04/06/2020)