PREDIÇÃO DE PREÇO DE VAREJO DE CERVEJA ARTESANAL

SSC0277 - Competições de Ciências de Dados

Nome: Victor Kendi Arakaki N°USP: 11219092

Sumário

1. Descrição do problema

A loja de cervejas artesanais russa possui dados dos produtos (cervejas) que vendem, bem como o histórico de transações ao longo do período de funcionamento. Foi proposto que a partir desses dados, verificar se o preço de varejo sugerido pelos próximos fornecedores é justo ou compatível com os padrões existentes já pelas cervejas artesanais vendidas no bar.

2. Análise dos dados

2.1 Explicação das variáveis

O repositório possui 2 tabelas de dados:

2.1.1 Product_range

Uma linha de produtos é o conjunto de todos os tipos e espécies de produtos (cerveja, petiscos, refrigerantes e etc.) oferecidos aos clientes pelo Bar Nelson Sauvin. Também pode ser entendido como um conjunto de produtos oferecidos por toda a indústria. Esta gama pode ser mais ou menos especializada ou genérica. É descrita pelo tamanho, teor alcoólico, unidade, preço, etc. A gama de produtos oferecidos deve corresponder às expectativas do mercado-alvo da empresa.

- Product_code: chave para merge com Transactions
    - Numérico
    - 5314 valores únicos
- Vendor_code: nome do fabricante
    - Categórico
    - 217 valores únicos
- Name: SKU
    - Categórica
    - 5193 valores únicos
- **Retail_price**: Preço de catálogo
    - Numérico
    - 184 valores únicos
- Base_unit: unidade do produto
    - Categórico
    - 4 valores únicos
- Country_of_Origin: país de origem
    - Categórico
    - 28 valores únicos
- Size: tamanho do item (SKU)
    - Numérico
    - 31 valores únicos
- ABV: Volume de álcool
    - Numérico

    - 151 valores únicos

2.1.2 Transactions

- Date_and_time_of_unloading: dae e hora quando o gerente abriu a caixa de registro
    - Categórico
    - 458 valores únicos
- Product_code: chave do produto para merge
    - Categórico
    - 3956 valores únicos
- Amount: número de unidades vendidas
    - Numérico
    - 617 valores únicos
- Sale_amount: total de dinheiro do montante que o negócio tem ganhado pela venda.
    - Numérico
    - 22572 valores únicos
- Discount_amount: é o dinheiro do montante que é deduzido do preço original do produto.
    - Numérico
    - 17843 valores únicos
- Profit: é a diferença entre a receita que uma empresa ganha e o custo associado a produção e venda.
    - Numérico
    - 27941 valores únicos
- Percentage_markup :é o montante pelo qual o custo de um produto é aumentado a fim de determinar o preço de venda.
    - Numérico
    - 13925 valores únicos
- Discount_percentage: é um desconto que é dado a um produto ou serviço que é dado como um montante por cem.
    - Numérico
    - 3629 valores únicos

A variável target do problema é a **Retail_price**, enquanto os outros atributos são utilizados na predição.

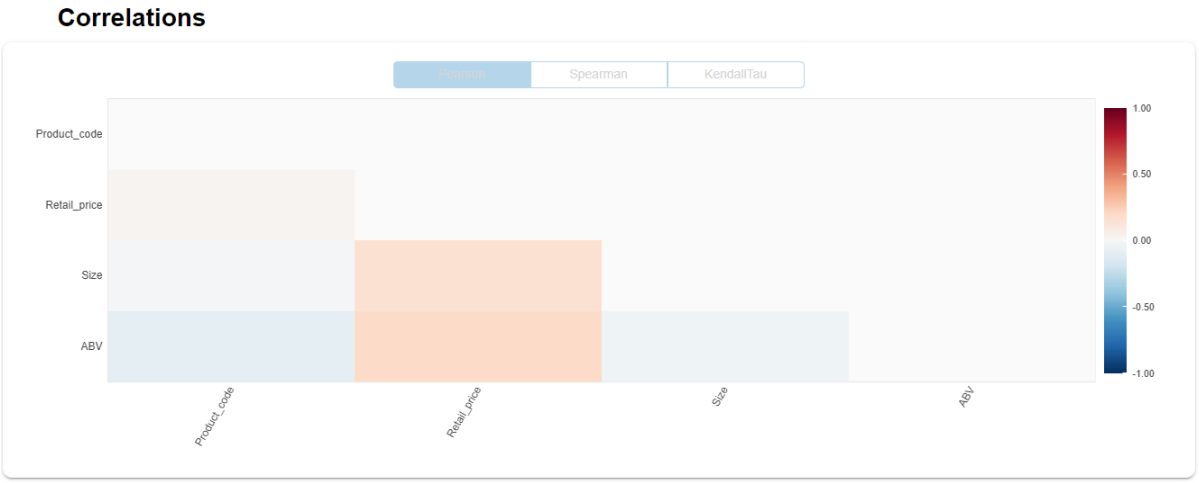
2.2 Análise dos dados

Utilizando a biblioteca DataPrep com o segmento de EDA (Exploratory data analysis), gerou-se um Report com estatísticas e insights sobre os dados do dataset.

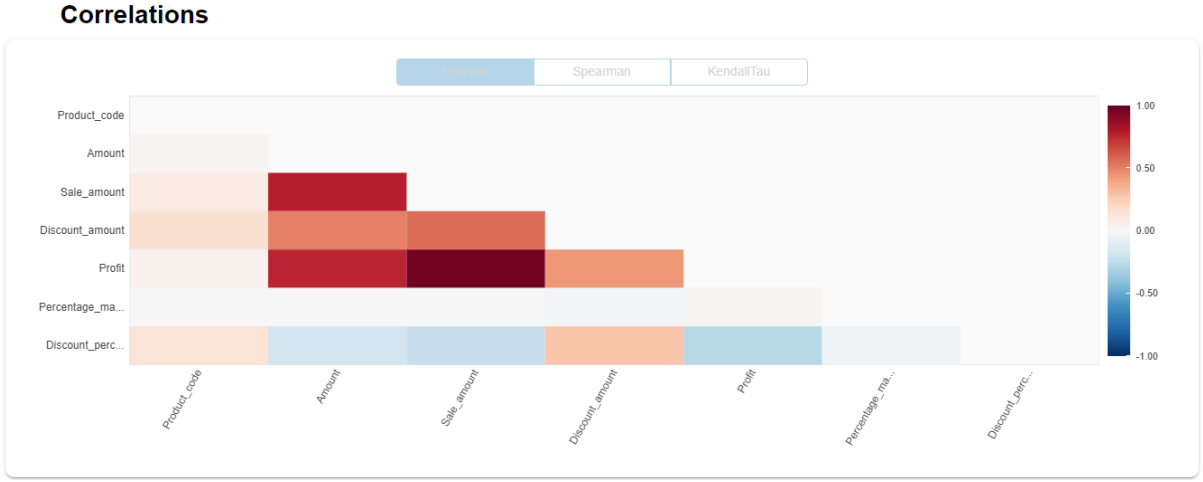
É verificável que o dataset possui 2 tabelas com 8 colunas cada e com a maioria das colunas com dados faltantes. A tabela Product_range possui os atributos Vendor_code, Retail_price, Base_unit, Country_of_Origin, Size e ABV com valores faltantes enquanto que a tabela Transactions possui os campos Sale_amount, Discount_amount, Profit, Percentage_markup e Discount_percentage. Para os valores numéricos, foi utilizado a estratégia de preencher os valores NaN com a média de cada campo, pois como o dataset

não possui muitos dados, optou-se por manter a maior quantidade de linhas. Já o atributo Vendor_code não passou por nenhum procedimento, pois não será utilizado. O campo Country_of_Origin, além de possuir dados faltantes, há relativa alta cardinalidade (28 valores únicos) e optou-se por agrupar os valores que não são “Russia” e preencher os valores NULOS com o valor “Unknown”. Os atributos relacionado ao montante e quantidade do produto como Retail_price, Size, Amount, Sale_amount, entre outros possuem muitos valores outliers bem maiores do que a média, mas não foram tratados, pois acredita-se que não são valores errados. A variável ABV parece seguir uma distribuição normal.

Abaixo, há a matriz de correlação de Pearson da tabela Product_range. Pode-se observar que não há nenhuma indício de forte correlação entre os atributos.



A seguir está a matriz de correlação de Pearson da tabela Transactions. Nota-se uma correlação muito forte entre os atributos Sale_Amount e Amount, Amount e Profit, Sale_Amount e Profit. Essas correlações são bem claras, pois são atributos podendo-se dizer dependentes um do outro. Optou-se por manter todos os atributos.



2.3 Benchmarks

As métricas que serão utilizadas para verificar a performance de cada algoritmo de classificação são:

- Erro Absoluto Médio (MAE)::

O MAE é uma métrica de avaliação de modelos de regressão que calcula a média das diferenças absolutas entre as previsões e os valores reais. Diferentemente do MSE, o MAE não eleva ao quadrado as diferenças entre as previsões e os valores reais. Isso significa que o MAE trata todos os erros de forma igual, independentemente de sua direção (positiva ou negativa). Por isso, o MAE é menos influenciado por valores extremos ou outliers, pois eles não são amplificados pelo quadrado.

- Coeficiente de Determinação (R^2):

O Coeficiente de Determinação é uma métrica amplamente utilizada para avaliar o desempenho de modelos de regressão. Ele fornece uma medida da proporção da variabilidade dos valores da variável dependente que é explicada pelo modelo.
O R^2 varia de 0 a 1, onde:
$R^2 = 0$ significa que o modelo não explica nenhuma variabilidade nos dados e as previsões são equivalentes à média dos valores observados.
$R^2 = 1$ significa que o modelo explica toda a variabilidade nos dados e faz previsões perfeitas.

3. Descrição e resultados das técnicas utilizadas

3.1 Técnicas para pré-processamento do dataset

Para lidar com a variável categórica Country_of_Origin, escolheu-se utilizar a técnica de LabelEncoder. Esta técnica é usada para codificar valores de uma única coluna em um conjunto de dados. Ele atribui um número inteiro exclusivo a cada categoria presente na coluna, começando de 0 para a primeira categoria, 1 para a segunda categoria, e assim por diante.

É notório ressaltar a aplicação da estratégia de preenchimento de dados faltante citados anteriormente utilizando a média no preenchimento dos atributos numéricos e a estratégia de agrupar os valores que não são “Russia” e preencher os valores NULOS com o valor “Unknown”

A junção das duas tabelas foi feito a partir de um código em SQL que faz JOIN pela chave Product_code e agrupa as transações para cada produto. Além disso, utiliza-se funções de agregação nos campos numéricos; para os dados de Product_range, utilizou-se AVG para manter os mesmos valores. Já para a tabela Transactions, a agregação de SUM foi utilizada nos campos de montate e quantidade para retornar o total, enquanto que AVG foi utilizado nos campos de porcentagem. Ademais, gerou-se um novo campo chamado Total_transacion que é a contagem do total de transações ocorridas para cada determinada cerveja artesanal vendida no bar.

O dataset foi separado em conjunto de teste e conjunto de treino na proporção de 0.2 e 0.8 utilizando a função train_test_split do sklearn com parâmetro shuffle=true.

3.2 Técnicas de Regressão

A seguir uma descrição sobre as técnicas de classificação utilizadas e resultados obtidos:

3.2.1 Regressão Linear Múltipla

- Regressão Linear Múltipla é um algoritmo de aprendizado de máquina que visa prever um valor contínuo com base em múltiplas variáveis independentes. Ele utiliza uma combinação linear dessas variáveis para estimar o valor da variável dependente, assumindo uma relação linear entre elas.

Métrica	**Regressão Linear Múltipla**
MAE	238.63
R^2	0.17

3.2.2 Regressão Ridge

- Regressão Ridge é uma técnica de aprendizado de máquina que se baseia na regressão linear, mas com uma penalização adicional nos coeficientes do modelo. Essa penalização ajuda a reduzir o efeito de multicolinearidade, quando as variáveis independentes estão altamente correlacionadas entre si. A regressão Ridge é especialmente útil quando se lida com conjuntos de dados com alta dimensionalidade e pode evitar o overfitting, melhorando a generalização do modelo.

Métrica	**Regressão Ridge**
MAE	238.63
R^2	0.17

3.2.3 Regressão Lasso

- Regressão Lasso é uma técnica de aprendizado de máquina semelhante à regressão Ridge, mas que utiliza uma penalização diferente nos coeficientes do modelo. Ao contrário da regressão Ridge, a regressão Lasso tem a capacidade de realizar seleção automática de variáveis, levando a um modelo mais esparso, onde alguns coeficientes podem ser exatamente zero. Isso torna a regressão Lasso útil para seleção de recursos e redução de dimensionalidade.

Métrica	Regressão Lasso
MAE	238.63
R^2	0.17

3.2.4 Regressão ElasticNet

- Regressão ElasticNet é uma combinação da regressão Ridge e da regressão Lasso. Ela utiliza tanto a penalização L1 (Lasso) quanto a penalização L2 (Ridge) nos coeficientes do modelo. A regressão ElasticNet visa obter os benefícios de ambas as técnicas, lidando com multicolinearidade, realizando seleção de variáveis e produzindo um modelo mais estável e robusto.

Métrica	Regressão ElasticNet
MAE	238.63
R^2	0.17

3.2.5 Máquina de Vetor de Suporte para Regressão

- Máquina de Vetor de Suporte para Regressão (SVR) é um algoritmo de aprendizado de máquina que visa prever um valor contínuo com base em um conjunto de variáveis independentes. Ele busca encontrar um hiperplano no espaço de atributos que melhor se ajusta aos dados de treinamento. A SVR é capaz de lidar com dados não lineares, através do uso de funções de kernel, e é amplamente utilizada em problemas de regressão, como previsão de preços, análise de séries temporais e modelagem estatística.

Métrica	Máquina de Vetor de Suporte para Regressão
MAE	664.86
R^2	-50.26

3.2.6 Árvores de Regressão

- Árvores de Regressão são algoritmos de aprendizado de máquina que dividem o espaço de atributos em regiões retangulares, com base em uma série de regras de decisão. Cada região representa uma previsão de valor contínuo para a variável dependente. A árvore de regressão é construída de forma recursiva, dividindo o espaço de atributos com base nas variáveis independentes e nos valores ótimos que melhor separam as instâncias. Elas são fáceis de interpretar e podem lidar com dados não lineares, além de serem utilizadas em várias áreas, como finanças, medicina e marketing.

Métrica (max depth = MAX)	**Árvores de Regressão**
MAE	96.35
R^2	0.65

3.2.7 Bagging Regressão

- Bagging Regressão é um algoritmo de aprendizado de máquina que utiliza a técnica de bagging para melhorar a acurácia de Regressão. Ele cria um conjunto de regressores independentes, cada um deles treinado em um subconjunto aleatório do conjunto de dados. A técnica é simples e pode ser utilizada em problemas de classificação e regressão, além de apresentar baixo risco de overfitting.
  O algoritmo Bagging sem a reposição de elementos é chamado de Pasting, o qual foi testado também para verificar se era gerado resultados melhores

Bagging Ensemble

Métrica	**Bagging Regressão**
MAE	252.36
R^2	-0.04

Pasting Ensemble

Métrica	**Pasting Ensemble**
MAE	252.36
R^2	-0.04

3.2.8 Random Forest Regressão

- Random Forest Regressão é um algoritmo de aprendizado de máquina que utiliza várias árvores de decisão para minimizar o erro na regressão. Ele cria uma floresta de árvores de decisão, cada uma delas treinada em um subconjunto aleatório do conjunto de dados. A técnica é robusta e pode ser utilizada em problemas de classificação e regressão, além de apresentar baixo risco de overfitting.

Métrica	**Random Forest Regressão**
MAE	128.88
R^2	0.71

3.2.9 AdaBoost Regressão

- O AdaBoost Regressão é baseado na ideia de combinar vários modelos de regressão fracos para formar um modelo mais robusto e preciso. Inicialmente, cada instância de treinamento tem um peso associado a ela, e um modelo de regressão fraco é ajustado a esses dados ponderados. Em seguida, os pesos são atualizados, dando mais importância às alterações que foram previstas incorretamente pelo modelo anterior.
  O algoritmo continua iterando, ajustando modelos fracos sucessivamente, até que um número pré-definido de iterações seja atingido ou até que um critério de parada seja satisfeito. Durante cada iteração, os modelos fracos são combinados usando uma média ponderada para gerar a previsão final.
  O AdaBoost Regressão tem a vantagem de lidar bem com dados complexos e não lineares. Ele é capaz de ajustar-se a estruturas de dados complicadas, fornecendo uma regressão precisa mesmo em cenários desafiadores. Além disso, o algoritmo permite a seleção automática de recursos, priorizando as variáveis mais relevantes para a tarefa de regressão.

Métrica	**AdaBoost Regressão**
MAE	339.43
R^2	0.11

3.2.10 XGBoost

- Utiliza árvores de decisão como modelos fracos e faz um ajuste iterativo dos pesos para melhorar a precisão das previsões. Durante cada iteração, o XGBoost ajusta uma nova árvore de decisão aos resíduos (diferenças entre os valores reais e as previsões atuais) do modelo anterior. O XGBoost Regressão também incorpora algumas técnicas avançadas para melhorar o desempenho e evitar overfitting, como a regularização L1 e L2 nos pesos das árvores, a limitação da profundidade das árvores e a amostragem estocástica das instâncias de treinamento. Uma das principais vantagens do XGBoost Regressão é a sua capacidade de lidar com conjuntos de dados grandes e complexos. Ele é eficiente em termos computacionais e pode lidar com milhões de instâncias e características. Além disso, o XGBoost possui mecanismos para lidar com valores ausentes e possui recursos incorporados para selecionar automaticamente as melhores características.

Métrica	**Regressão Linear Múltipla**
MAE	54.17
R^2	0.93

3.3 Descrição dos resultados obtidos

Observa-se nos resultados obtidos que as técnicas mais simples que regressão linear tiveram péssimos desempenhos para esta tarefa, o que pode ter sido influenciado pelos campos com alta correlação que não foram excluídos por opção. Mesmo assim, não se espera um bom resultado. A partir da árvore de decisão de regressão, começa-se a observar resultados mais satisfatórios, com MAE de 96.35 e R^2 score de 0.65, o que não chega a ser um resultado tão bom. Já o random forest a alcançou um valor maior de MAE de 128.58, mas um R^2 de 0.71, podendo se equiparar com a árvore de decisão. A técnica escolhida para superar o baseline foi a XGBRegressor, o qual alcançou um resultado muito acima do esperado, com MAE de 54.17 e um R^2 de incríveis 0.93, o que coloca em outro patamar de modelo para a resolução deste problema. Observa-se que esta ferramenta poderosa é muito superior que as apresentadas anteriormente.

4. Conclusão

O resultado obtido leva a crer que o XGBRegressor será sempre superior a este problema e possivelmente para problemas mais complexos, já que este também é um modelo mais complexo. Para problemas mais simples, talvez não compense o custo

computacional e tempo de processamento que levaria para utilizar esta poderosa técnica, podendo ser considerado o teste das técnicas de regressão linear e as inspiradas em árvore que obtiveram melhores resultados no problema atual e são menos pesadas do que o XGBRegressor.

5. Referências

Código desenvolvido:
https://colab.research.google.com/drive/1PnV8nUWXFEIN0Folh43iK4seLK6vB7P_?usp=sharing
Dataset:
https://www.kaggle.com/datasets/podsyp/sales-in-craft-beer-bar

6. Apêndice

Resumo das técnicas de classificação apresentadas no dia 14/07:

6.1 Modal Linear Regression

Victor Gomes apresentou sobre um novo modelo de regressão denominado Regressão Linear Modal, que foi concebido para explorar dados de elevada dimensão. O modelo estima a modalidade de uma distribuição com base em estimadores de densidade de kernel não paramétricos e modela a modalidade condicional de uma variável de resposta dado um conjunto de fatores de previsão como uma função linear desses fatores de previsão. Explicou também que este modelo oferece várias vantagens em relação à regressão linear padrão, incluindo maior precisão e robustez na presença de valores atípicos. Além disso, percebe-se pode ser utilizada para modelar relações não lineares entre os fatores de previsão e a variável de resposta.

6.2 Predictive modeling of blood pressure during hemodialysis: a comparison of linear model, random forest, support vector regression, XGBoost, LASSO regression and ensemble method

O artigo apresenta um estudo sobre modelagem preditiva da pressão arterial durante a hemodiálise. O objetivo do estudo foi desenvolver um sistema inteligente para a criação de perfis e previsão da pressão arterial durante a hemodiálise, com o objetivo de evitar a hipotensão intradialítica (IDH) e melhorar os resultados dos pacientes.

Para atingir esse objetivo, os pesquisadores compararam seis métodos diferentes de modelagem preditiva: modelos lineares, random forest, regressão de vetor de suporte, XGBoost, regressão LASSO e métodos de conjunto. Eles coletaram dados sobre a pressão arterial e a frequência de pulso dos pacientes durante a hemodiálise usando o dispositivo de gateway Vital Info Portal (VIP) e o sistema de saúde digital. Eles também coletaram outras

informações clínicas, como dados demográficos, comorbidades, tipo de acesso vascular, relação cardiotorácica, medicação cardíaca, Kt/V, frequência de IDH e exames laboratoriais.

O estudo constatou que os algoritmos de aprendizado de máquina podem prever com eficácia as alterações na pressão arterial durante a hemodiálise. O método de conjunto foi considerado o método mais preciso para prever alterações na pressão arterial sistólica durante a hemodiálise. As descobertas têm implicações importantes para melhorar os resultados dos pacientes durante a hemodiálise, prevenindo a IDH.

6.3 Estimating the change in soccer's home advantage during the Covid-19 pandemic using bivariate Poisson regression

O artigo discute como a pandemia de Covid-19 afetou os jogos de futebol em todo o mundo, com muitos jogos sendo adiados e eventualmente remarcados para serem disputados sem a presença de torcedores. Os autores argumentam que isso representa uma oportunidade de estudar o impacto de estádios vazios sobre a vantagem de jogar em casa no futebol.

Para analisar esse impacto, os autores usam modelos de regressão de Poisson bivariados, que, segundo eles, são mais apropriados do que os modelos de regressão linear para esse tipo de dados. Eles demonstram, por meio de simulações, que a regressão de Poisson bivariada reduz em quase 85% o viés absoluto ao estimar o benefício da vantagem de jogar em casa em uma única temporada de jogos de futebol.

Usando dados de 17 ligas de futebol profissional, os autores usam modelos de Poisson bivariados para estimar a mudança na vantagem do mandante devido a jogos disputados sem torcedores. Eles concluíram que há uma redução significativa na vantagem de jogar em casa quando os jogos são disputados sem torcedores, com algumas ligas apresentando reduções maiores do que outras.