Previsão de Consumo de Cerveja Utilizando Machine Learning

Lucas da Silva Passos, Edson Kazumi Yamamoto, Fernando Ferreira Duarte, Yuri Lucas de Lima

Engenharia da Computação – Faculdade de Engenharia de Sorocaba (FACENS)
Sorocaba – SP – Brasil

Lucas.spassos@outlook.com, edsonkazyamamoto@gmail.com,
feernando.duarte@gmail.com, yurilucaslima17@gmail.com

Abstract. *When we talk about commerce, the first thing we think is what is the correct quantity to be acquired so that there is no lack of product for customers and not enough to create waste. Predicting the amount to be sold is very important so that you do not have lost sales and do not spend more than necessary, often keeping perishable products in stock. Looking at this problem, we have identified the opportunity to create a model that can analyze a historical sales base and create a learning using regression that makes this forecast more closely based on historical consumption.*

Resumo. *Ao falar em comércio, logo nos vem a cabeça, qual a quantidade correta a ser adquirida para que não falte produto para os clientes e não sobre o suficiente para criar desperdícios. Prever a quantia a ser vendida é muito importante para que não tenha vendas perdidas e não gaste mais que o necessário, mantendo muitas vezes, produtos perecíveis em estoque. Ao observar esse problema, foi identificada a oportunidade de criar um modelo que consiga analisar uma base histórica de vendas e criar um aprendizado utilizando regressão que faça essa previsão de forma mais aproximada baseado em um consumo histórico.*

1. Introdução

Hoje a tomada de decisão baseada em sentimentos tem se tornado uma ação opcional em algumas áreas. A área do comércio tem se tornado um grande utilizador das tecnologias para tomada de decisão. Devido a quantidade de variáveis presentes em

nosso dia a dia, fica difícil de analisá-las e chegar a uma conclusão assertiva. Prever a quantidade de um produto, principalmente perecível, a ser comprado na medida certa é totalmente desafiador. Se um produto é comprado demais, pode acontecer desse produto sobrar, ser danificado ou perder a validade e perder a possibilidade da venda. Já quando o produto é comprado em falta, clientes podem ficar sem produto, desistir da compra e até mesmo conhecer outros lugares e não voltarem mais. Tudo isso devido a falta de assertividade na escolha do estoque, que é feita a partir de uma escolha emocional. Se tiver uma forma de conciliar os dados passados numa base histórica, conciliando variáveis que tenham ligação com o tipo de produto e isso tudo se tornasse um modelo matemático que nos informasse um valor muito próximo do que deveria ser, aumentaria o lucro do estabelecimento e a satisfação dos clientes. A proposta apresentada é um aprendizado de máquina que utiliza dados históricos de um estabelecimento que vende cerveja, uma das bebidas mais consumidas no mundo. Os dados (amostra) foram coletados em São Paulo - Brasil, em uma área universitária, onde existem alguns happy hours, com grupos de jovens de 18 a 28 anos de idade. O conjunto de dados utilizado possui 7 atributos, sendo um alvo, com período de um ano de coleta. O alvo no nosso caso é estimar a quantidade em litros de cerveja que será consumida.

2. Problema encontrado

2.1. Inteligência Artificial no Comércio

A procura das empresas por soluções digitais tem se tornado cada dia mais comum, visto que existe uma grande aderência de todos os públicos por tecnologias, como celulares, tablets ou smartwatchs. Algumas soluções voltadas ao comércio, tem se tornado extremamente eficiente, mostrando relatos de comerciantes que chegaram a aumentar suas vendas em até 70%.

Mesmo tendo algumas tecnologias muito bem difundidas para essa área, foi percebida a falta de uma tecnologia que pudesse auxiliar os comerciantes de alimentos perecíveis na tomada de decisão. Como saber qual a quantidade de alimentos a serem adquiridos para um final de semana? Como vai ser o movimento de um feriado? Investir ou não? Essas são perguntas extremamente frequentes desse ramo. Um estoque deve ser muito bem preciso para que não sobre alimentos e os mesmos precisem ser descartados por possuírem baixo tempo de consumo.

2.2. Especialidade escolhida

A partir do problema encontrado, o ramo escolhido foi o de bebidas. Já que existe uma grande procura dos jovens por happy hours após o trabalho, ou em dias de jogos de futebol. Segundo o site Cidade de São Paulo, existem 30 mil bares na cidade. Isso sem contar os 20 mil restaurantes, que podem oferecer serviços parecidos.

Com a inteligência artificial, o problema de saber o quanto investir para um final de semana, ou feriado, passa ser muito mais assertiva. Seria apenas necessário criar uma base histórica de consumo do principal item, no caso, a cerveja, aliado a algumas informações como o dia da semana, se esse dia é feriado, ou também o clima.

2.3. Coleta de dados

Foi estudado algumas bases históricas, onde encontramos uma base de dados constituída de informações coletadas durante um ano. Os dados (amostra) foram coletados em São Paulo, em uma área universitária, onde existem alguns happy hours, com grupos de jovens de 18 a 28 anos de idade. O conjunto de dados utilizado possui 7 atributos, sendo um alvo, o consumo de cerveja em litros. Com período de um ano de coleta.

3. Técnica de predição e ferramentas

Com o dado devidamente estruturado, a linguagem de programação mais indicada para a análise dos dados foi Python. O intuito da aplicação é predizer a quantidade de cerveja que será vendida num determinado dia do ano. O princípio de funcionamento é carregar a base histórica, que está armazenada num arquivo separado por vírgula (CSV). Esse dado passará por uma etapa de análise, para que possa verificar a existência de outliers e também a distribuição dos dados. Nessa etapa utiliza-se algumas bibliotecas do Python, o seaborn, biblioteca utilizada para criar gráficos estatísticos, e o matplotlib que é uma biblioteca de plotagem para a linguagem de programação Python e sua extensão de matemática numérica NumPy. Ele fornece uma API orientada a objetos para incorporar gráficos em aplicativos. Então será passado por uma etapa de pré-processamento para ajustar aos dados para que não haja impacto negativo durante o aprendizado. Como precisamos chegar a um valor contínuo, dentre as técnicas apresentadas em sala, será utilizada a Regressão linear.

4. Análise do dataset e pré processamento

O Dataset é composto pelas colunas, Data, Temperatura Média (C), Temperatura Mínima (C), Temperatura Máxima (C), Precipitação (mm), Final de Semana (Booleano), Consumo de cerveja (litros). Após os dados serem carregados, foi utilizado a função Head para observar como estava a qualidade dos primeiros dados da amostra. Isso pode nos mostrar uma pequena parte dos dados e verificarmos como esses dados foram preenchidos (Figura 1).

	Data	Temperatura Media (C)	Temperatura Minima (C)	Temperatura Maxima (C)	Precipitacao (mm)	Final de Semana	Consumo de cerveja (litros)
0	2015-01-01	27,3	23,9	32,5	0	0	25.461
1	2015-01-02	27,02	24,5	33,5	0	0	28.972
2	2015-01-03	24,82	22,4	29,9	0	1	30.814
3	2015-01-04	23,98	21,5	28,6	1,2	1	29.799
4	2015-01-05	23,82	21	28,3	0	0	28.900

Figura 1. Comando HEAD

Apenas com esses dados já é possível tomar algumas decisões, como modificar os dados para que não haja vírgula separando os números. Isso fará com que os dados não sejam entendidos como texto e não fará com que o aprendizado “entenda” errado esses atributos.

Em seguida foi removida a data da coleta. Esse dado não contribui para o aprendizado do algoritmo, podendo criar um o*verfitting* , como demonstrado na figura 2. O o*verfitting* ocorre quando o modelo se ajusta aos dados, ou seja, o modelo serve só para os dados da base que foi utilizada para a sua construção. O que ocorre é que nesse caso o modelo passa em diversos testes de precisão com o conjunto de dados utilizados, porém, não serve para predição. Em outras palavras, como alguns cientistas de dados costumam dizer, o seu modelo aprende os dados da base treino ao invés de aprender o todo e ser capaz de fazer previsões.

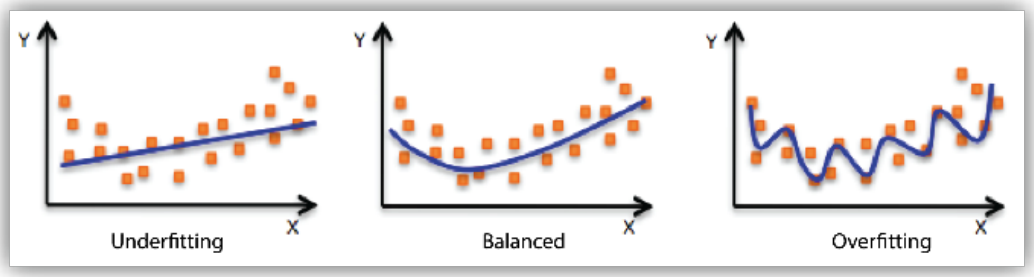

Figura 2. Ajuste dos dados

Agora é possível analisar de uma forma geral para enxergar como está a distribuição dos dados. Existem diversas formas de analisar a distribuição dos dados. A primeira e mais simples que foi utilizada foi o método “describe”. Esse método traz algumas informações do dataset, como máximo, mínimo e média de cada classe, assim como a distribuição em quartis. A Figura 3 mostra o resultado dessa função no dataset estudado.

	Temperatura Media (C)	Temperatura Minima (C)	Temperatura Maxima (C)	Precipitacao (mm)	Final de Semana	Consumo de cerveja (litros)
count	365.000000	365.000000	365.000000	365.000000	365.000000	365.000000
mean	21.226356	17.461370	26.611507	5.196712	0.284932	25.401367
std	3.180108	2.826185	4.317366	12.417844	0.452001	4.399143
min	12.900000	10.600000	14.500000	0.000000	0.000000	14.343000
25%	19.020000	15.300000	23.800000	0.000000	0.000000	22.008000
50%	21.380000	17.900000	26.900000	0.000000	0.000000	24.867000
75%	23.280000	19.600000	29.400000	3.200000	1.000000	28.631000
max	28.860000	24.500000	36.500000	94.800000	1.000000	37.937000

Figura 3. Distribuição através do describe

Com essa função podemos ver se existem outliers na amostra, como dados com valor zero, que podem não combinar com o atributo, como os valores de temperatura já que estamos falando da cidade de São Paulo, que dificilmente chegaria a temperaturas menores que zero. Também podem ser observado os valores máximos vendo se as temperaturas não passam de valores fora da realidade para esses locais. Em seguida utilizamos o pairplot (figura 4) da biblioteca seaborn. Aqui é possível ver a correlação entre os atributos.

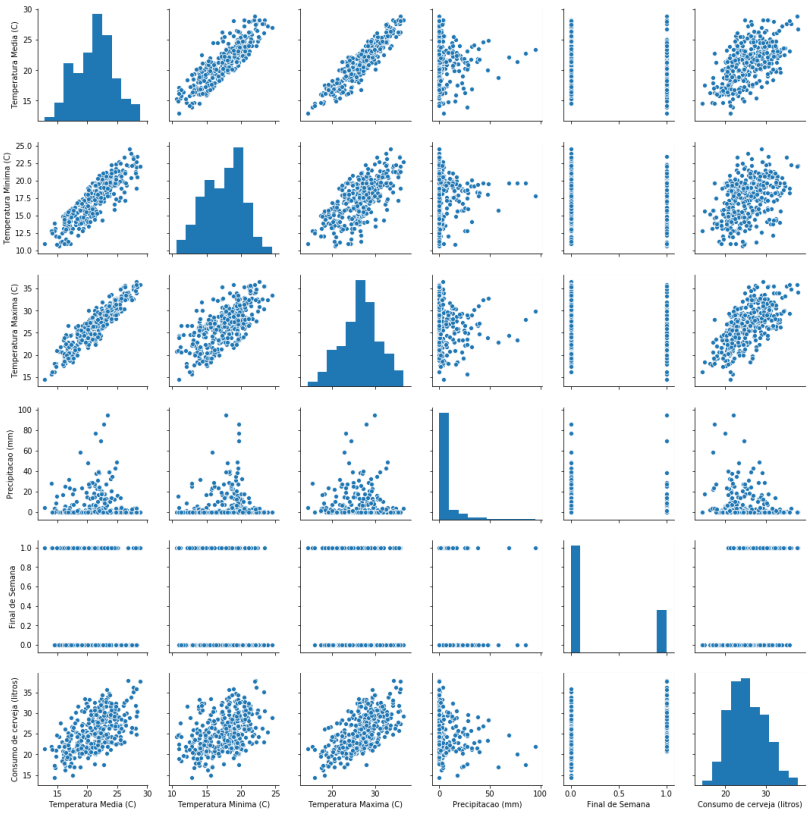

Figura 4 . Pairplot

7. Regressão Linear

Análise de regressão é uma metodologia estatística que utiliza a relação entre duas ou mais variáveis quantitativas de tal forma que uma variável possa ser predita a partir de outra. O objetivo da regressão linear é buscar a equação de uma linha de regressão que minimize a soma dos erros ao quadrado, da diferença entre o valor observado de Y e o valor previsto.

8. Resultados

Após analisarmos os gráficos, definimos que as colunas que mais são características do problemas são as de Temperatura Máxima (C) e Final de Semana, realizamos o train_test_split e treinamos nosso modelo para regressão linear com 70% dos dados.

Utilizando algumas das várias métricas disponíveis, como Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) e Mean absolute percentage error (MAPE), executamos uma predição com apenas um dado e comparamos com um dado que tínhamos conhecimento, obtivemos os seguintes resultados:

	MAE	MSE	RMSE	MAPE	MAPE com validação cruzada
Valores de erro	0.19166	0.11020	0.33197	0.62201%	8.14% (+/- 1.37)

Tabela 1 - Tabela de erros

O MAPE com validação cruzada é o resultado mais confiável, já que ele embaralha os dados e seu resultado é em porcentagem, totalizando 8.14% +/- 1.37, o que é um resultado aceitável.

O modelo gerou a seguinte fórmula: y = 5.4 + 0.7 * Temperatura Máxima + 5.1 * Final de semana, aproximadamente. O que significa que para cada C° a previsão de consumo de cerveja aumenta em 0,7L e se for final de semana aumenta em 5,1L, e por padrão somamos 5,4L no dia.

Em seguida realizamos outro train_test_split desta vez com 30% de dados para treino, veja os novos resultados na tabela 2.

	MAE	MSE	RMSE	MAPE	MAPE com validação cruzada
Valores de erro	0.25433	0.19405	0.44051	0.82538%	8.14% (+/- 1.37)

Tabela 2 - Tabela de erros 2

O valor com validação cruzada manteve-se apesar de tudo, já que ele usa uma separação própria, porém os outros parâmetros tiveram um leve aumento no erro, mas considerando que está usando apenas 30% da base de dados para treino é plausível considerar essa solução.

Agora mostraremos os erros totais, usando todos os dados que foram usados no treino, com os que foram separados para o teste, e obtivemos os seguintes resultados:

	MAE	MSE	RMSE	MAPE	MAPE com validação cruzada
Valores de erro	1.93240	5.26760	2.29512	7.78162%	8.14% (+/- 1.37)

Tabela 3 - Tabela de erros da base completa

Comparando o MAPE sem e com a validação cruzada, os valores são semelhantes. Os erros se encontra em uma margem adequada, já que os valores de MAE e RMSE estão próximo de 2, tanto na tabela 3 quanto na tabela 4.

A tabela 3 está para os valores de 70% de treino e 30% de teste, enquanto a tabela 4 está com 30% treino e 70% teste, comparando ambas as tabelas observa-se que a diferença entre os valores é mínima, sendo assim viável a utilização de menos dados para o modelo, melhorando sua performance.

	MAE	MSE	RMSE	MAPE	MAPE com validação cruzada
Valores de erro	2.02379	6.03037	2.45568	8.10588%	8.14% (+/- 1.37)

Tabela 4 - Tabela de erros da base completa

9. Conclusão

Com regressão conseguimos fazer previsões para o futuro e apoiar decisões de negócio e até ter novos Insights dos dados que não sabíamos que estavam ali. É possível termos tecnologias como essa nos ajudando no dia a dia e facilitando a tomada de decisões muitas vezes complexas. Mas não podemos esquecer de tratar os dados de forma que o nosso modelo não fique mal modelado, contendo outliers, ou coisas fora do padrão, que não identifiquem o que está sendo estudado.

Sem dúvidas o comércio está tomando um novo rumo com o uso dessas tecnologias. Hoje tudo é informação. A todo momento estamos fornecendo informações, principalmente comportamentais, que podem ser usadas como base na construção de tecnologias como essa. Cada dia, estamos mais imerso no mundo digital e é inevitável que tecnologias como essa, se ajustem mais e mais a modelos matemáticos que podem

facilitar todo nosso trabalho “invisível” ou de alta complexidade com uma exatidão muito maior.

10. Referências Bibliográficas

URFGS. (2019). “Álgebra Linear - Regressão Linear Simples”.
ufrgs.br/reamat/AlgebraLinear/livro/s14-regressx00e3o_linear_simples.html

São Paulo, Cidade (2019). “Por que São Paulo?”.
http://cidadedesaopaulo.com/v2/pqsp/dados-e-fatos/?lang=pt