SCC0277 - Competições de Ciências de Dados

Desafio: Regressão

Aluno: Diego Giaretta de Paulo

Universidade de São Paulo (USP)

Venda de cervejas artesanais

Dados sobre as vendas de cervejas artesanais em um bar russo foi disponibilizado para que seja feita uma análise de regressão, com a variável resposta podendo ser elegida pelo autor. Sendo assim, neste trabalho a variável escolhida foi o percentual de markup, ou seja, a quociente entre o valor de venda e os custos envolvidos na transação.

Os dados possuem uma grande quantidade de valores vazios que serão tratados ao longo do problema.

Engenharia de dados

	Date_and_time_of_unloading	Product_code	Amount	Sale_amount	Discount_amount	Profit	Percentage_markup	Discount_percentage	Vendor_cod
0	2020-01-01 23:00:00	144	1.0	280.00	NaN	155.00	124.00	NaN	AF Bre
1	2020-01-01 23:00:00	209	2.0	545.73	294.27	75.73	16.11	35.03	Pohja
2	2020-01-01 23:00:00	213	2.0	1265.05	34.95	653.05	106.71	2.69	Kerisa
3	2020-01-01 23:00:00	217	1.0	630.00	70.00	220.50	53.85	10.00	Savo
4	2020-01-01 23:00:00	222	2.0	1104.75	195.25	393.75	55.38	15.02	Bello

Como é possível observar, estes dados apresentam uma mescla entre as informações de transações e também as informações de cada produto transacionado. É notável que dentre as informações tem-se o valor da venda, do lucro e do preço de varejo, as porcentagens de markup e desconto e descritivos das cervejas como ABV, país de origem, marca e nome da cervjea.

	Vendor_code	Retail_price	Base_unit	Country_of_Origin	Size	ABV	target
0	AF Brew	280.0	Pieces	Russia	0.330	10.3	1.2400
1	Pohjala	420.0	Pieces	Estonia	0.330	6.8	0.1611
2	Kerisac	650.0	Pieces	France	1.000	6.0	1.0671
3	Savoie	870.0	Pieces	France	0.750	4.5	0.5385
4	Bellot	770.0	Pieces	France	0.750	5.0	0.5538
5	Boon	540.0	Pieces	Belgium	0.375	7.0	0.0047
6	Verhaeghe	900.0	Pieces	Belgium	0.750	6.2	0.2008
7	Founders	420.0	Pieces	USA	0.355	5.7	0.0775
8	BrewDog	540.0	Pieces	United Kingdom	0.660	3.8	0.1987
9	Schneider Weisse	370.0	Pieces	Germany	0.500	8.2	1.4737

Após a limpeza dos dados, obtivemos estas features para o modelo. Sendo a porcentagem de markup a variável de interesse neste problema.

```
Quantidade de dados faltantes
```

```
Vendor_code          6324
Retail_price            3
Base_unit               0
Country_of_Origin    8507
Size                 5793
ABV                  5807
target               1939
dtype: int64
```

Como pode-se notar, o conjunto de dados conta com um valor expressivo de valores ausentes, inclusive na nossa variável de interesse 'target'. Para isso, todos as instâncias sem a porcentagem de markup serão removidas dos dados do modelo. Para lidar com os valores ausentes nas features numéricas (ABV, size e retail_price), as medianas das respectivas colunas serão utilizadas para substituir os dados ausentes.

Para lidar com as váriaveis categóricas, a abardogem escolhida foi a criação de váriaveis dummies, ou seja, cada categoria dentro de país de origem, marca e medida receberá uma coluna booleana (verdadeiro ou falso) indicando se aquela **instâ**ncia é ou não daquela categoria. Para cada país e medida foi criada uma variável dummy e somente para as 10 marcas de cervejas com mais produtos vendidos.

Modelagem

Para a modelagem, os dados foram divididos em 80% para treino e 20% para teste.

Para a construção de uma baseline, os modelos de regressão linear, KNN e SVM foram selecionados e a seguir são disponibilizados os seus resultados.

```
Linear Regression Mean squared error: 203.33
Linear Regression R Score: 0.00
KNN Mean squared error: 361.90
KNN R Score: -0.77
SVM Mean squared error: 204.11
SVM R Score: -0.00
```

Como é possível notar, pela baseline criada, será necessário uma melhor coleta de dados e buscar novas abordagens. Um modelo CatBoost será implementado para verificar uma possível melhora dos resultados, mas é notável que estas váriaveis ainda são pouco relevantes para a resolução deste problema.

```
CatBoost Mean squared error: 200.85
CatBoost R Score: 0.02
```

Com resultados também não animadores, para uma melhora este vendedor de cervejas artesanais deverá passar por uma nova coleta de dados e buscar uma nova abordagem, por exemplo uma abordagem no contexto de séries temporais, aproveitando os dados hisóricos indicados na análise exploratória de dados.

Conclusão

Neste gráfico é possível enxergar a influência das váriaveis no modelo, indicando um desempenho razoável apenas das váriaveis numéricas, principalmente o preço de varejo. Podemos notar que apenas cervejas apresentam um bom desempenho na importância das váriaveis.

```
<Axes: >
```