Universidade do Minho
Especialização em *Business Intelligence*
Unidade Curricular de Análise de Dados

Ano Letivo de 2016/2017
2º Semestre

Análise de Dados de Histórico de Vendas e Inventário Ativo com recurso a metodologia CRISP-DM

Carlos Sá – A59905
João Lopes – A61077
Grupo 4

Abril, 2017

Data de Recepção	
Responsável	
Avaliação	
Observações	

Análise de Dados de Histórico de Vendas e Inventário Ativo com recurso a metodologia CRISP-DM

Carlos Sá – A59905
João Lopes – A61077
Grupo 4
Abril, 2017

Resumo

O presente documento constitui o resultado de um projecto de Análise de Dados no contexto da especialização em Business Intelligence do Mestrado Integrado em Engenharia Informática da Universidade do Minho. Este projeto conta com uma análise de um conjunto de dados sobre *Histórico de Vendas e Inventário Ativo*. Esta análise será feita de acordo com metodologia adequada e utilizada na área dos processos de aprendizagem automática que foram discutidos na unidade curricular de Análise de Dados da presente especialização. A metodologia de análise utilizada neste projeto será a metodologia *CRISP-DM* amplamente utilizada na área de Data Mining.

O conjunto de dados a analisar fazem parte de um dataset obtido na plataforma Kaggle. O mesmo possui 14 atributos e aproximadamente 200 000 casos que caracterizam o histórico de vendas e de inventário atual para venda de uma entidade.

É efetuado um levantamento de questões consideradas relevantes para análise, a partir de uma análise inicial dos atributos bem como da prévia compreensão do modelo de negócio da entidade. Após isto, é seguida a metodologia CRISP-DM para dar resposta às questões de análise consideradas, utilizando processos de aprendizagem automática.

Área de Aplicação: Data Mining - Análise de Dados com aplicação de metodologias de aprendizagem automática com recurso a metodologia CRISP-DM.

Palavras-Chave: Data, mining, análise, dados, crisp-dm, business, modelos, linguagem R, R-Studio, modelo logístico, árvores de decisão, random forests, machine learning, aprendizagem automática.

Índice

Índice de Figuras

Índice de Tabelas

1. Introdução

Na área do *Data Mining* a análise de dados e extração de informação é um dos dos instrumentos mais importantes no processo de tomada de decisão. O termo *Data Mining* é uma área da Engenharia Informática enquadrada na atividade de Business Intelligence que organiza, procura padrões, associações, anomalias entre outras informações relevantes de um determinado conjunto de dados (*dataset*). Num projeto de data mining existe um conjunto de etapas que se conseguem distinguir relativamente à exploração dos dados, construção de um modelo, procura de padrões, validação e verificação. Diferentes técnicas de recuperação de dados e inteligência artificial podem ser aplicadas com objetivo de encontrar padrões nos dados, correlações e estatística que permitam extrair conhecimento com relevância para uma entidade.

O trabalho pelo qual este relatório recai tem como objetivo a análise de um *dataset* de uma entidade cujo negócio é a venda de produtos. Esse *dataset* contém informação acerca de produtos que foram vendidos nos últimos 6 meses anteriores à disponibilização do mesmo e ainda dos produtos que se encontram atualmente em inventário para venda.

Pretende-se que seja efetuada uma análise inicial aos atributos do dataset por forma a perceber o significado de cada um bem como as suas estatísticas. Após esta análise inicial deve ser definido um conjunto de questões de análise às quais deve ser dada resposta a partir de métodos de aprendizagem automática (*machine learning*) tirando partido da linguagem *R*.

Por forma a dar resposta às questões de análise bem como à perceção correta dos dados contido no *dataset*, é seguida a metodologia *CRISP-DM* amplamente utilizada na área de *Data Mining*. Começa-se assim por descrever o contexto de operação da entidade que disponibilizou o *dataset* e definição das várias questões de análise passando depois para uma análise inicial dos vários atributos (i.e. descrição, domínios, estatísticas) que permite selecionar à priori atributos que sejam menos relevantes para análise. Após esta análise inicial é efetuada uma análise mais completa dos atributos que inclui métodos como análise de correlação entre os mesmos e ainda uma análise da qualidade dos dados do *dataset*.

Uma vez realizadas as tarefas acima descritas é efetuada a preparação dos dados, onde são selecionados os dados de treino e de teste, passando para a modelação onde são aplicadas várias técnicas de aprendizagem automática sobre os dados por forma a dar resposta à questões de análise definidas.

O ultimo passo de todo este processo é, naturalmente, a avaliação dos resultados obtidos dos modelos de aprendizagem desenvolvidos.

2. Motivação

Com o recurso a ferramentas de apoio à decisão, como o data mining, numa determinada entidade surge a oportunidade de a partir de um conjunto de dados potencialmente elevado, extrair conhecimento que permita direcionar os esforços de uma entidade num sentido que lhe permita maximizar lucros, expansão do negócio para novos clientes, otimizar vendas de um determinado produto entre outros. A aquisição desse conhecimento é muitas vezes feita recorrendo a técnicas sofisticadas de exploração de dados, análises estatísticas e correlações. Técnicas como o Data Mining, quando utilizadas, permitem transformar clientes eventuais em clientes fieis, criar estratégias de vendas mais adequadas entre outras.

O segredo é saber como os produtos e serviços devem ser vendidos direcionando as vendas para o cliente certo, na altura certa com o preço adequado. Num contexto empresarial o data mining representa uma grande oportunidade de explorar dados de vendas e dos seus clientes por forma a prever boas oportunidades de compra, criar estratégias diferenciadas de venda que melhorem o ciclo de vida da empresa. Este trabalho prático, reflete a importância e a oportunidade da utilização de técnicas de exploração de dados em *Business Intelligence* como o *Data Mining*. Neste é disponibilizado um conjunto de dados considerável com cerca de 200 000 registos relativo a um histórico de vendas e inventário ativo que pertence a uma determinada entidade.

A análise de um dataset realista permite que se obtenha experiencia sobre a forma como os dados são apresentados no dia-a-dia que, em muitos casos, podem ser grandes quantidades de dados e de qualidade questionável para o objetivo que é necessário atingir com os mesmos.

Pretende-se então que se tenha um contato mais direto com um conjunto de dados próximo do real por forma a realizar uma analise inicial para aprofundamento dos dados e levantamento de um conjunto de questões relevantes para análise que são passíveis de serem respondidas com a exploração de técnicas de mineração de dados sobre todos os registos do *dataset* a analisar. Ao longo deste trabalho, o data mining será utilizado com o recurso a linguagem *R* para explorar os dados, efetuar cruzamento estatístico de informação relevante e construção de modelos que permitam dar resposta a esse conjunto de questões.

3. Compreensão do Negócio

A compreensão do negócio (*Business Understanding*) é o primeiro passo a seguir no processo de desenvolvimento de análise de um conjunto de dados. Pretende-se que, antes de efetuar qualquer tipo de análise aos dados referentes ao negócio, fique claro os motivos e objetivos que levam ao desenvolvimento do processo de análise de dados bem como ter claramente definidas todas as infraestruturas necessárias para o sucesso de todo o processo. Estas podem incluir todo um conjunto de recursos necessários como disponibilização monetária adequada para a longevidade da análise dos dados, acesso ao conjunto dos dados necessários e relevantes para análise e ainda a determinação de riscos inerentes.

Uma vez que o caso de estudo presente é puramente académico, apenas vai ser efetuada uma breve descrição da entidade de negócio a partir da informação disponibilizada e a definição do conjunto de questões de análise que se pretendem modelar.

3.1. Entidade de negócio

O negócio em análise é referente a uma entidade que possui um *dataset* cujos dados representam as vendas dos seus produtos. Estes dados estão divididos em histórico de vendas dos últimos 6 meses e inventário ativo presente atualmente.

A entidade possui uma grande quantidade de produtos no seu inventário e apenas uma pequena parte dos mesmos são vendidos bem como muitos deles são vendidos apenas uma vez no decorrer de um ano.

Com a implementação de um modelo de análise de dados pretende-se que os dados referentes ao inventário ativo presente sejam alvo de uma análise recorrendo a processos de aprendizagem automática por forma a ser possível prever a melhor forma de atuação sobre os produtos em inventário.

3.2. Levantamento de questões de análise

Através da observação dos dados contidos no *dataset* e depois de uma análise preliminar aos mesmos, com o objetivo de perceber que tipo de informação é possível obter, foi possível elaborar um conjunto de questões de análise às quais se pretende dar resposta e assim ajudar a entidade a atuar de forma mais apropriada no futuro. A análise de todas as questões é efetuada recorrendo a métodos de aprendizagem automática.

É então proposta a análise de:

1. Quais as variáveis mais relevantes para análise com o objetivo de obter um resultado mais preciso possível para cada uma das restantes questões de análise.

Isto leva a que tenha de ser efetuada uma análise detalhada a cada uma das variáveis.

2. Probabilidade de venda de cada produto (consoante o conjunto de variáveis escolhido e o melhor modelo obtido – logístico, árvores de decisão ou *random forests*).
3. Quantidade esperada de venda de cada produto (de acordo com o melhor modelo conseguido).
4. Quantidade apropriada de inventário para cada produto onde, possivelmente, quantidades iguais a 0 ou inferiores a um certo valor devem/podem ser interpretadas como produtos que devem ser descontinuados.
5. Qual o valor monetário que é esperado obter na venda de produtos em inventário. Pretende-se que seja efetuada uma previsão do preço de venda em conjunto com uma previsão de venda dos produtos (inclui uma análise aos 3 tipos de preços presentes no *dataset* por forma a perceber qual o melhor a ser utilizado que pode ser efetuada na questão 1 ou na presente questão).

Ao longo das próximas etapas de análise vão ser apresentadas formas de dar resposta a cada uma das questões de análise propostas bem como fundamentar a escolha destas questões com os dados presentes no *dataset*.

4. Compreensão dos Dados

Esta fase de desenvolvimento de um modelo de análise de dados é caracterizada pela exploração mais detalhada dos dados para análise. É utilizado uma série de métodos de exploração de dados por forma a perceber o seu tamanho, numero de atributos, tipo e variação dos atributos, qualidade dos dados, entre outros.

Os dados para análise foram obtidos numa plataforma web – *Kaggle*[1] – e contêm um conjunto de informação acerca do histórico de vendas de produtos de uma entidade nos últimos 6 meses e ainda informação acerca dos produtos que possuem atualmente em inventário para venda.

Os resultados de toda a análise efetuada sobre os dados foram obtidos pela utilização da linguagem *R*. Qualquer outro tipo de linguagem ou software utilizado ao longo da análise dos dados será feita a sua referência aquando da sua utilização.

4.1. Análise e descrição dos atributos

Considere-se a tabela 1 com o extrato dos primeiros produtos contidos no *dataset*:

Order	File_Type	SKU_number	SoldFlag	SoldCount	Marketing Type	Release Number	New_Release_Flag	Strength Factor	PriceReg	Release Year	Item Count	LowUser Price	LowNet Price
2	Historical	1737127	0	0	D	15	1	682743	44.99	2015	8	28.97	31.84
3	Historical	3255963	0	0	D	7	1	1016014	24.81	2005	39	0.00	15.54
4	Historical	612701	0	0	D	0	0	340464	46.00	2013	34	30.19	27.97
6	Historical	115883	1	1	D	4	1	334011	100.00	2006	20	133.93	83.15
7	Historical	863939	1	1	D	2	1	1287938	121.95	2010	28	4.00	23.99
8	Historical	214948	0	0	D	0	0	1783153	132.00	2011	33	138.98	13.64

Tabela 1: Extrato dos primeiros produtos do *dataset*

Order	FileType	SKUnumber	SoldFlag	SoldCount	Marketing Type	Release Number	NewReleaseFlag	Strength Factor	PriceReg	Release Year	Item Count	LowUser Price	LowNet Price

Tabela 2: Nome dos atributos do *dataset* depois de efetuar uma remoção do caractere "_"

Linhas	198917
Colunas	14

Tabela 3: Número de linhas e colunas/atributos do dataset

A tabela 3 mostra que o *dataset* possui cerca de 200.000 linhas e 14 atributos. Na tabela 2 é possível observar o nome dos atributos depois de efetuar a remoção do caractere "_" do nome dos atributos que o possuíam por forma a que todos possuam um nome de formato similar.

[1] https://www.kaggle.com/flenderson/sales-analysis

A tabela abaixo contém uma descrição de cada atributo e do seu formato para que seja possível perceber mais detalhadamente o que cada atributo representa.

Atributo	Descrição	Formato
Order	Contador sequencial dos registos.	Int
FileType	Tipo de registo. "Historical" se o registo pertencer ao histórico de vendas ou "Active" se o registo pertencer ao conjunto de produtos em inventário para venda.	String
SKUnumber	Identificador unívoco de um produto. O mesmo produto pode ocorrer no histórico de vendas como no inventário ativo.	Int
SoldFlag	Identificador binário que indica se ocorreu ou não uma venda nos últimos 6 meses: "0" caso não tenha ocorrido uma venda e "1" caso tenha ocorrido pelo menos uma venda. Este identificador é aplicado apenas para os registos históricos. Para os registos em inventário ativo não é apresentada qualquer informação acerca da ocorrência de venda sendo este atributo representado por "NA".	Binário/String
SoldCount	Representa a quantidade vendida de cada produto nos últimos 6 meses. É, portanto, maior ou igual a "SoldFlag". É aplicado apenas para os registos históricos. Para os registos em inventário ativo não é apresentada qualquer informação acerca da quantidade vendida (uma vez que também não apresentada informação acerca da ocorrência de uma venda) sendo este atributo representado por "NA".	Int/String
MarketingType	Tipo de marketing do produto. Dois tipos de marketing: "D" e "S". Devem ser considerados independentes um do outro.	Char
ReleaseNumber	Identificador do lançamento. Provavelmente não acrescenta qualquer tipo de informação relevante.	Int
NewReleaseFlag	Identifica se um produto teve anteriormente um novo lançamento. "0" em caso negativo e "1" caso tenha tido um novo lançamento.	Binário
StrengthFactor	Valor representativo da resistência do produto.	Int
ReleaseYear	Ano de lançamento do produto.	Int
ItemCount	Quantidade em inventário de cada produto.	Int
PriceReg	Diferentes tipos de preços de venda. Não foi possível obter qualquer informação acerca da distinção entre eles.	Float
LowUserPrice		
LowNetPrice		

Tabela 4: Descrição e formato dos atributos do *dataset*

No anexo 1 pode ser consultada informação adicional, acerca de cada um dos atributos, obtida pela utilização das funções *str()* e *summary()* da linguagem *R*.

A descrição dos atributos permite que seja possível obter uma primeira sensibilidade acerca dos atributos mais relevantes para análise e aqueles que não fornecem qualquer tipo de informação adicional ou relevante.

Atributos considerados irrelevantes e que poderão ser excluídos da análise são:

- Order, SKUnumber e ReleaseNumber

Uma vez que os dados dizem respeito a vendas de produtos, os atributos considerados mais relevantes para uma primeira análise são:

- FileType, SoldFlag e MarketingType

4.1.1 Análise inicial de atributos considerados relevantes :: FileType, SoldFlag e MarketingType

Numa primeira análise aos dados é necessário perceber as suas estatísticas. Uma vez que o *dataset* contém dados históricos e dados de inventário ativo, torna-se evidente a necessidade de perceber a distribuição dos mesmos uma vez que vão ser, provavelmente, utilizados para a divisão do *dataset* em dados de treino e de teste na fase de preparação de dados.

Active	122921
Historical	75996

Tabela 5: Quantidade de registos do tipo "Historical" e "Active"

Active	61.79512
Historical	38.20488

Tabela 6: Percentagem de registos do tipo "Historical" e "Active"

Como é possível observar pelas tabelas 5 e 6, existem 75996 registos do tipo histórico, que representam 38.2% dos registos do *dataset* e 122921 registos de produtos em inventário ativo, que representam 61.8% dos registos do *dataset*.

A variável "SoldFlag" vai ser, muito provavelmente, a variável com maior relevância para análise bem como para o processo de aprendizagem automático. Para tal, é necessário perceber os seus domínios em cada tipo de registo.

	0	1
Active	0	0
Historical	63000	12996

Tabela 7: Quantidade de produtos de cada tipo que não foram vendidos (0) e que foram vendidos(1)

	0	1
Active	0.0000	0.0000
Historical	82.8991	17.1009

Tabela 8: Percentagem de produtos de cada tipo que não foram vendidos (0) e que foram vendidos(1)

As tabelas 7 e 8 mostram, respetivamente, a quantidade e percentagem de produtos históricos e ativos nos casos onde ocorreu ou não uma venda. Os dados de inventário ativo encontram-se a zero uma vez que o *dataset* não possui informação acerca da venda deste tipo de dados. Nos dados históricos, 63000 são produtos que não foram vendidos nos últimos 6

meses e apenas 12996 produtos foram vendidos nesse intervalo de tempo. Isto representa uma percentagem de 82.9 de produtos não vendidos para 17.1% de produtos vendidos.

Desta forma, é possível observar um desbalanceamento dos dados deste atributo que, uma vez que é um atributo binário e que é o atributo chave de análise bem como de aprendizagem, este desbalanceamento pode levar a que o método de aprendizagem utilizado seja influenciado por este desbalanceamento, tendendo para o caso de maior percentagem. No processo de modelação, por forma a verificar se tal está de facto a acontecer, é necessário olhar para além do valor de precisão (*accuracy*) que embora possa ser elevado, os valores de "precision", "recall" e "f-score" são mais precisos na avaliação do modelo. No caso de o processo de aprendizagem ser influenciado por este desbalanceamento vai ser necessário tratar o mesmo recorrendo a um método adequado como por exemplo a reamostragem do *dataset* (*resampling dataset*).

Uma vez que existem duas formas de marketing dos produtos, numa primeira análise, pode ser relevante perceber qual a distribuição dos dois tipos de marketing. A tabela abaixo mostra a quantidade de produtos com cada tipo de marketing. Como é possível observar, existe uma distribuição praticamente uniforme de produtos em cada tipo de marketing.

D	97971
S	100946

Tabela 9: Quantidade de produtos com cada tipo de marketing

Embora exista praticamente tantos produtos em ambos os tipos de marketing, é relevante inferir a distribuição dessas quantidades em cada tipo de registo, como mostra a tabela abaixo:

	ACTIVE	HISTORICAL
D	62852	35119
S	60069	40877

Tabela 10: Quantidade de produtos em cada tipo de marketing por cada tipo de registo

É possível observar que em cada tipo de registo existe aproximadamente a mesma quantidade de produtos dos dois tipos de marketing. Existe, portanto, um balanceamento quase perfeito deste atributo que leva a que seja um possível candidato para pertencer aos modelos de aprendizagem. Por forma a tomar essa decisão, irá ser efetuada, posteriormente, uma análise acerca de qual tipo de marketing possui mais influência sobre a venda de produtos.

4.1.2 Análise inicial dos restantes atributos

(A ser efetuada posteriormente)

4.1.3 Correlação entre atributos

Tendo em conta a quantidade de atributos e os seus significados no contexto do *dataset*, torna-se importante perceber qual a medida da relação entre cada um dos atributos por forma a encontrar aqueles que se correlacionam de forma significativa, para que possam ser alvo de uma análise mais detalhada.

Uma vez que os dados estão separados em dois tipos de registos (históricos e ativos), torna-se evidente que os dados alvo para teste no processo de aprendizagem são os ativos, logo, os dados históricos devem ser utilizados para treino. Isto leva à necessidade de separação dos dados em dois *datasets*, um com os dados históricos e outro com os dados de inventário ativo. Após a separação do *dataset* original foi necessário efetuar uma seleção dos atributos que devem fazer parte desta análise. Foram excluídos os atributos "*Order*", "*SKUnumber*" e "*ReleaseNumber*" uma vez que, como referido anteriormente, não trazem qualquer tipo de informação adicional ou relevante devido à sua natureza e descrição. Para além destes atributos foram ainda excluídos todos aqueles que não sejam numéricos: "*FileType*" e "*MarketingType*".

Uma vez efetuada esta seleção dos dados, foram aplicadas as funções "*pairs"* e "*cor*", da linguagem R, no subdataset de dados históricos, por forma a obter duas formas distintas de visualização da correlação entre as variáveis.

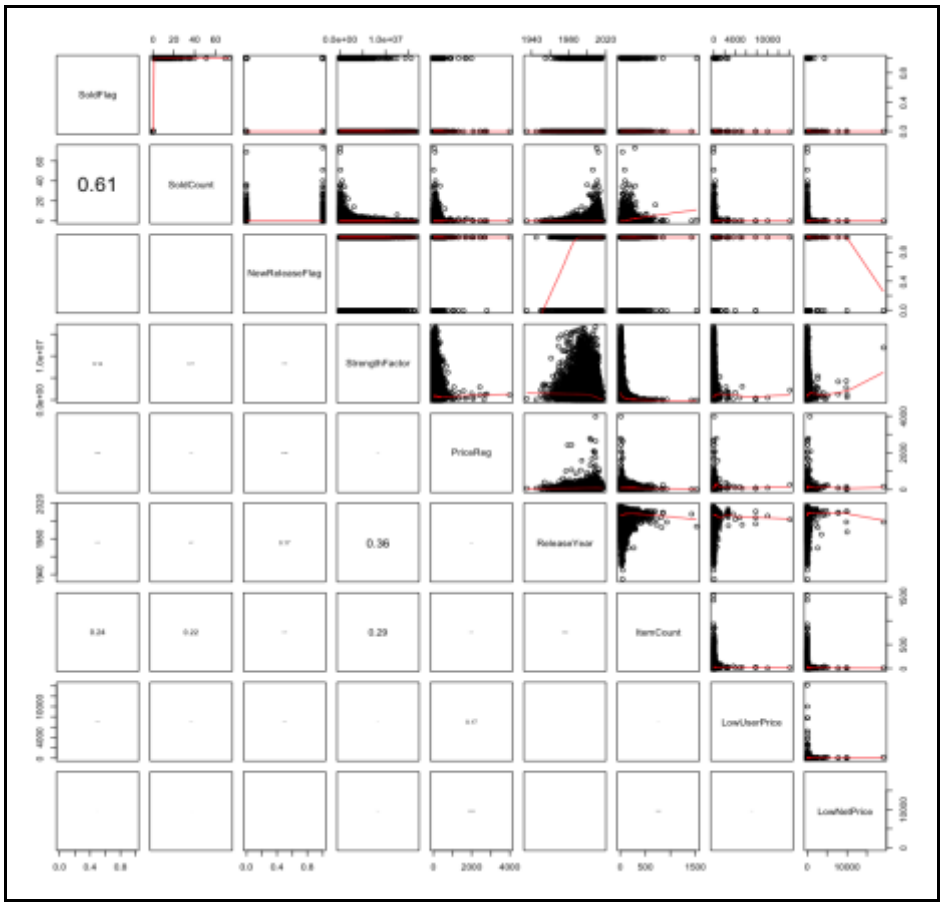

Figura 1: Output da função "*pairs*" sobre os dados históricos

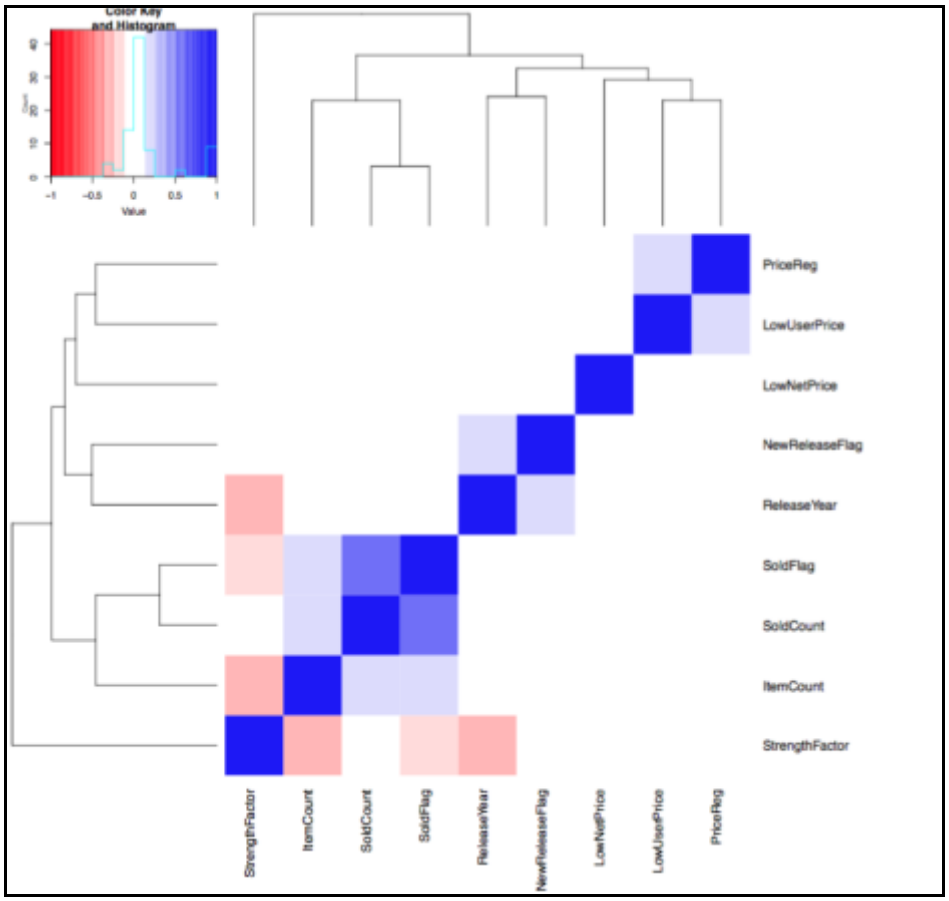

Figura 2: Output da função "*cor*" sobre os dados históricos em formato de *heatmap*

As figuras 1 e 2 apresentam os outputs das funções "*pairs*" e "*cor*" respetivamente. O output da função "*pairs*" apresenta na parte superior um conjunto de *scatterplots* e na parte inferior o valor numérico da correlação entre as variáveis. A função "*cor*" tem um output onde apenas os atributos correlacionados são apresentados como um *heatmap*.

Os dois *outputs* acima permitem observar que os atributos não possuem um grande corelacionamento entre eles. Posteriormente será efetuada uma análise sobre os atributos que possuem uma correlação entre si.

4.2. Qualidade dos dados

Após toda a análise efetuada anteriormente é também necessário perceber a qualidade dos dados presentes no *dataset*. Esta análise da qualidade dos dados é um fator importante uma vez que pode comprometer os resultados obtidos nos modelos a desenvolver.

Assim, após uma cuidada observação dos atributos bem como dos seus valores e estatísticas, não foram encontradas ocorrências de valores nulos (com exceção dos registos de inventário ativo que não possuem informação acerca da ocorrência de vendas/quantidades vendidas uma vez que são os dados a serem alvo de teste) que é um problema bem conhecido na análise de dados.

Contudo, embora não existam valores nulos, podem ser referidos alguns problemas com a qualidade da informação que o *dataset* oferece. Esses problemas passam por:

- Desbalanceamento do atributo "*SoldFlag*" que é o atributo fundamental de análise. Este desbalanceamento pode levar a que o modelo de aprendizagem escolhido seja influenciado.
- Falta de informação contextual sobre o significado de certos atributos como os vários tipos de preços bem como a moeda em que se encontram e escala do valor de resistência dos produtos.
- Falta de uma etiqueta temporal com informação acerca da data de venda de cada produto. A inclusão de tal etiqueta no *dataset* permitiria prever com maior certeza a venda de produtos num intervalo temporal específico e não apenas no futuro.

I. Anexo 1 – Análise Inicial dos Atributos :: *str* & *summary*

- *str()*

```
'data.frame':   198917 obs. of  14 variables:
 $ Order         : int  2 3 4 6 7 8 9 10 11 12 ...
 $ FileType      : Factor w/ 2 levels "Active","Historical": 2 2 2 2 2 2 2 2 2 2 ...
 $ SKUnumber     : int  1737127 3255963 612701 115883 863939 214948 484059 146401 110568 764270 ...
 $ SoldFlag      : int  0 0 0 1 1 0 0 0 0 0 ...
 $ SoldCount     : int  0 0 0 1 1 0 0 0 0 0 ...
 $ MarketingType : Factor w/ 2 levels "D","S": 1 1 1 1 1 1 1 1 1 1 ...
 $ ReleaseNumber : int  15 7 0 4 2 0 13 4 11 5 ...
 $ NewReleaseFlag: int  1 1 0 1 1 0 1 1 1 1 ...
 $ StrengthFactor: num  682743 1016014 340464 334011 1287938 ...
 $ PriceReg      : num  45 24.8 46 100 122 ...
 $ ReleaseYear   : int  2015 2005 2013 2006 2010 2011 2010 2011 2008 2004 ...
 $ ItemCount     : int  8 39 34 20 28 33 33 57 36 19 ...
 $ LowUserPrice  : num  29 0 30.2 133.9 4 ...
 $ LowNetPrice   : num  31.8 15.5 28 83.2 24 ...
```

Figura 3: Output da função "*str()*" sobre o *dataset*

- *summary()*

Order	FileType	SKUnumber	SoldFlag	SoldCount	MarketingType
Min. : 2	Active :122921	Min. : 50001	Min. :0.00	Min. : 0.00	D: 97971
1st Qu.: 55665	Historical: 75996	1st Qu.: 217252	1st Qu.:0.00	1st Qu.: 0.00	S:100946
Median :108569		Median : 612208	Median :0.00	Median : 0.00	
Mean :106484		Mean : 861363	Mean :0.17	Mean : 0.32	
3rd Qu.:158298		3rd Qu.: 904751	3rd Qu.:0.00	3rd Qu.: 0.00	
Max. :208027		Max. :3960788	Max. :1.00	Max. :73.00	
			NA's :122921	NA's :122921	

ReleaseNumber	NewReleaseFlag	StrengthFactor	PriceReg	ReleaseYear	ItemCount
Min. : 0.000	Min. :0.0000	Min. : 6	Min. : 0.00	Min. : 0	Min. : 0.00
1st Qu.: 1.000	1st Qu.:0.0000	1st Qu.: 161419	1st Qu.: 42.00	1st Qu.:2003	1st Qu.: 21.00
Median : 2.000	Median :1.0000	Median : 582224	Median : 69.95	Median :2007	Median : 32.00
Mean : 3.412	Mean :0.6422	Mean : 1117115	Mean : 90.89	Mean :2006	Mean : 41.43
3rd Qu.: 5.000	3rd Qu.:1.0000	3rd Qu.: 1430083	3rd Qu.: 116.00	3rd Qu.:2011	3rd Qu.: 50.00
Max. :99.000	Max. :1.0000	Max. :17384454	Max. :12671.48	Max. :2018	Max. :2542.00

LowUserPrice	LowNetPrice
Min. : 0.00	Min. : 0.00
1st Qu.: 4.91	1st Qu.: 17.95
Median : 16.08	Median : 33.98
Mean : 30.98	Mean : 46.83
3rd Qu.: 40.24	3rd Qu.: 55.49
Max. :14140.21	Max. :19138.79

Tabela 11: Output da função "*summary()*" sobre o *dataset*