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

Junho, 2017

bi

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
Junho, 2017

Indice

Anexos

Índice de Figuras

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.

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.

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

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 figura 1 abaixo apresenta um conjunto de gráficos (histogramas e barplots) dos restantes atributos dos dados históricos (que é o conjunto de dados alvo para a aprendizagem). Esta representação gráfica dos atributos permite obter uma visualização geral do domínio de cada um, permitindo obter um maior conhecimento de cada variável do dataset por forma a auxiliar na escolha das variáveis a serem utilizadas em cada modelo a ser implementado.

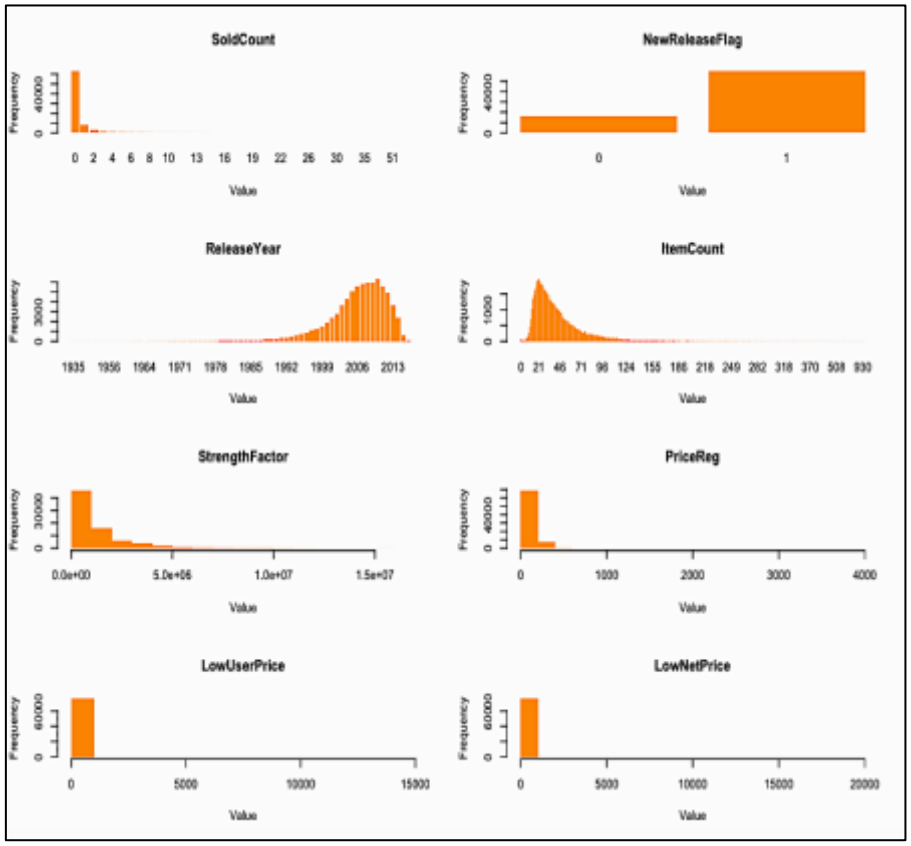

Figura 1: Representação gráfica das restantes variáveis

Para a variável *SoldCount* é possível observar que a maior parte dos produtos não são vendidos, como já foi visto na análise anteriormente efetuada à variável *SoldFlag*.

Em relação a novos lançamentos de produtos existentes, *NewReleaseFlag*, existem quase três vezes mais aqueles que tiveram um novo lançamento em relação aos restantes.

É possível observar uma distribuição, quase normal, nos anos de lançamento dos produtos, estando o pico de maior lançamento em torno do ano 2010.

Para os produtos em inventário, *ItemCount*, existe um valor relativamente disperso sendo o valor mínimo 0 e o máximo 1523 (valor não observável no gráfico, mas obtido manualmente). Contudo, é para valores próximos de 20 onde ocorre a maior frequência de produtos em inventário.

O gráfico dos valores representativos da resistência, *StrengthFactor*, mostra que existem mais produtos com um fator de resistência menor.

Por fim, para as variáveis representativas dos vários tipos de preços, *PriceReg, LowUserPrice e LowNetPrice*, é possível observar que a maior frequência de valores acontece em valores baixos de preços. Uma vez que não foi possível obter qualquer tipo de informação que distinga os vários tipos de preços, estas três variáveis vão ser alvo de uma análise mais detalhada por forma a compreender se existe alguma relação significativa entre elas e de que forma influenciam a potencialidade de vendas de produtos.

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.

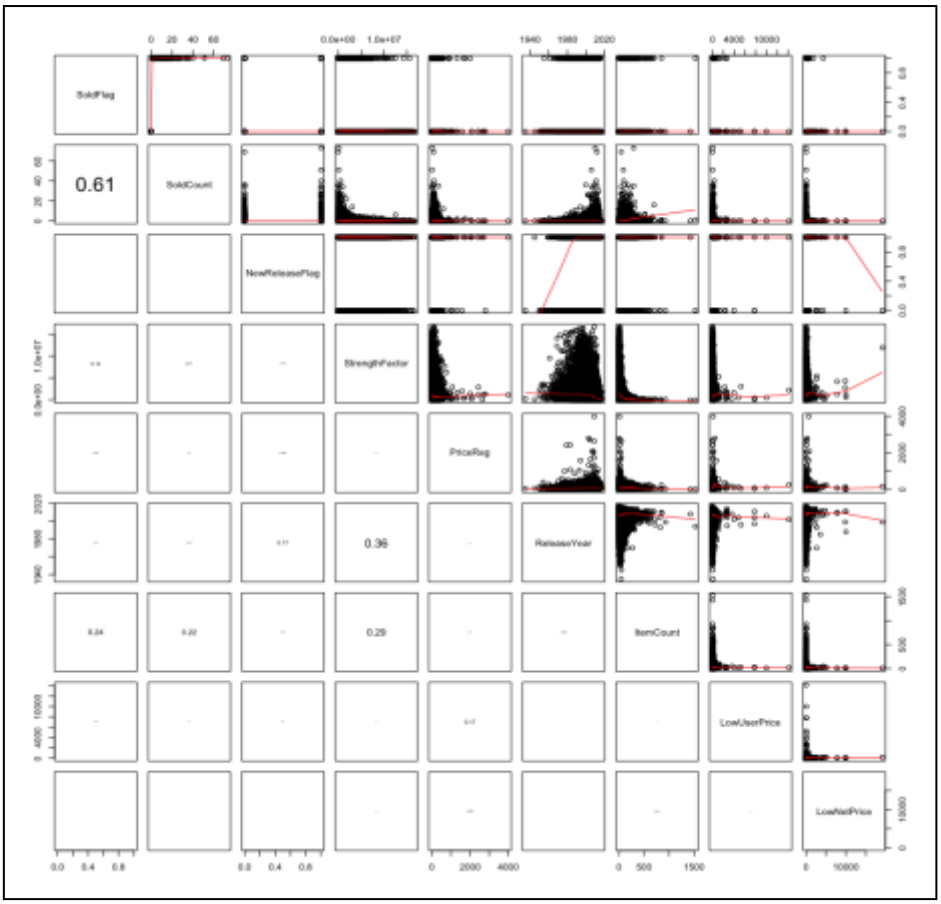

Figura 2: Output da função "*pairs*" sobre os dados históricos

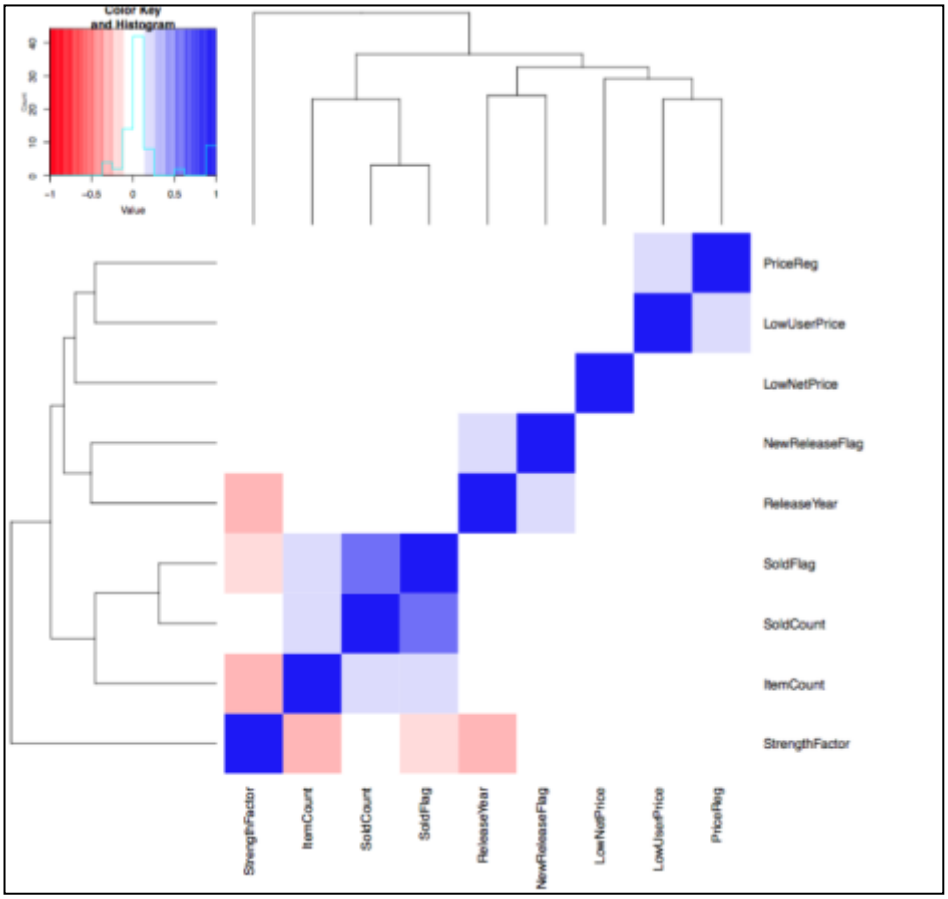

Figura 3: Output da função "*cor*" sobre os dados históricos em formato de *heatmap*

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 uma grande correlação 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.

5. Preparação dos Dados

Nesta fase de desenvolvimento vai ser efetuada uma análise mais detalhada às variáveis do *dataset*. Com esta análise procura-se compreender de forma mais detalhada o impacto que cada variável apresenta sobre a potencialidade de venda de um produto que é a análise fundamental a ser efetuada sobre o *dataset*.

Esta análise vai permitir que seja efetuada uma seleção mais apropriada das variáveis a serem incluídas em cada modelo a ser desenvolvido bem como facilitar a limpeza dos dados como a ocorrência de possíveis valores omissos, tratamento de *outliers*, entre outros.

5.1. Tratamento de *outliers*

A presença de valores atípicos nos dados, designados de *outliers*, que podem ocorrer por vários motivos como erros na coleta de dados ou eventos raros que não estão relacionados com o fenómeno que se quer aprender, podem prejudicar o processo de aprendizagem devido ao aumento de complexidade e ruido nos dados de treino. Quando tal acontece, deve ser efetuada uma análise por forma a compreender se esses dados devem ser tratados, que na maioria dos casos passa pela sua remoção do conjunto de dados de treino.

Foi efetuada uma análise à presença de *outliers* no conjunto de dados de análise. Desta análise foi possível observar que, de uma forma geral, todas as variáveis relevantes para análise contêm *outliers*, como pode ser observado na figura 4, que apresenta uma representação visual em formato de *boxplot* da variação de algumas das variáveis do *dataset*.

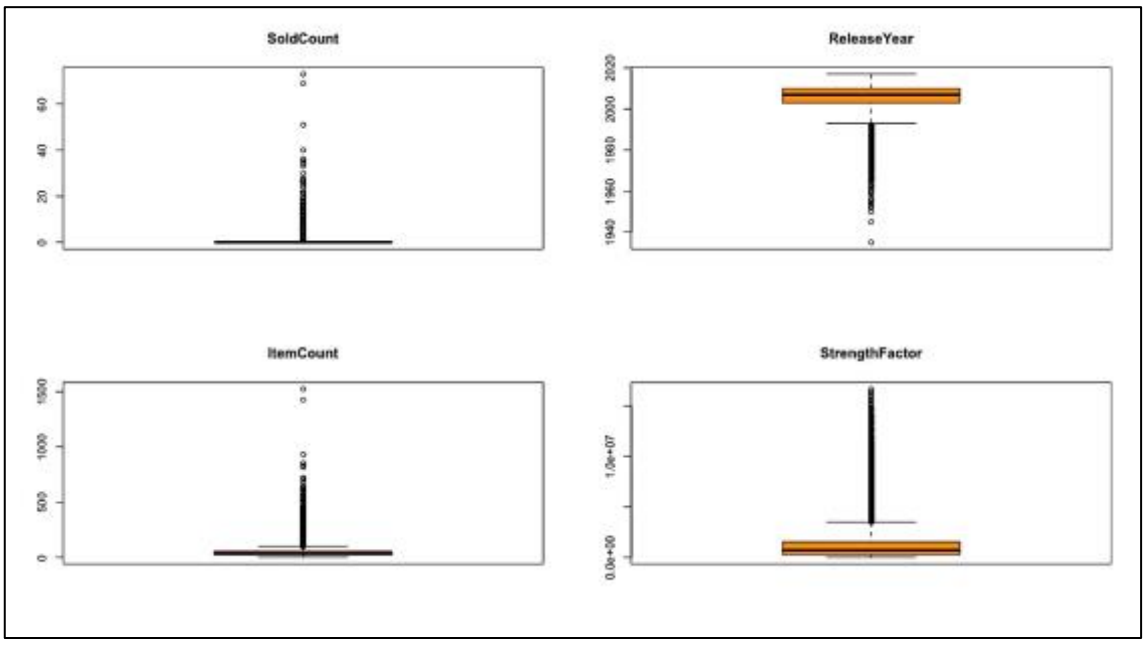

Figura 4: *Outliers* presentes nas variáveis *SoldCount, ReleaseYear, ItemCount* e *StrengthFactor* dos dados históricos

Dada a informação apresentada na figura 4, foi necessário decidir qual medida a tomar face à presença de *outliers* nas variáveis do *dataset*. Tendo em conta a natureza do negócio que o conjunto de dados representa, onde ocorrem um número reduzido de vendas, foi considerado que a remoção dos *outliers*, que em certas variáveis atinge os milhares (3000+) e tendo em conta o tamanho do *dataset* e considerando o numero reduzido de vendas de cada produto, onde a maior parte dos produtos não são vendidos, embora sejam *outliers*, representam uma grande quantidade de informação para que possa ser descartada, estando a ser perdida, potencialmente, informação de vendas que ocorreram e que são de grande importância para determinar a venda de um produto.

Desta forma, foi decidido que os *outliers* presentes no conjunto de dados não serão removidos por representarem uma grande quantidade de dados, evitando perder informação essencial acerca das vendas de produtos.

Esta decisão pode, no entanto, como referido anteriormente, trazer complexidade e ruído nos modelos de aprendizagem a serem desenvolvidos podendo levar a que, por exemplo, um certo modelo que necessite de dados o mais linear possíveis, não os tenha. Nestes casos, o modelo deve ser analisado e tratado para tentar reduzir ao máximo a não linearidade dos dados. Em alternativa, e como trabalho posterior, podem ser construídos modelos com o conjunto de dados sem *outliers* e efetuar a comparação da qualidade dos modelos (com e sem *outliers* nos dados).

5.2. Impacto das variáveis sobre a venda de produtos

Anteriormente foi analisada a variação de cada variável bem como a sua importância inicial para pertencerem aos modelos de aprendizagem. Isto permitiu que certas variáveis fossem previamente consideradas menos relevantes (*Order*, *SKUnumber* e *ReleaseNumber*) e mais relevantes.

Dada a natureza do conjunto de dados e do negócio que representam, a variável mais importante é a que determina a ocorrência de venda em cada produto (*SoldFlag*). Torna-se assim importante efetuar uma análise mais detalhada a cada uma das restantes variáveis. Mais precisamente, esta análise deve focar-se na observação do impacto que cada uma apresenta sobre a venda de produtos nos dados históricos.

Esta análise vai possibilitar que se possua um maior conhecimento da correlação entre as várias variáveis e a venda de produtos, permitindo potencialmente, selecionar apenas as variáveis que se considerem mais relevantes para as vendas por forma a serem utilizadas nos modelos de aprendizagem a serem desenvolvidos.

5.2.1 Impacto de *MarketingType* sobre *SoldFlag*

Existem dois tipos de *marketing* pelos quais os produtos são comercializados. A perceção do impacto que cada tipo apresenta sobre a venda de produtos torna-se relevante.

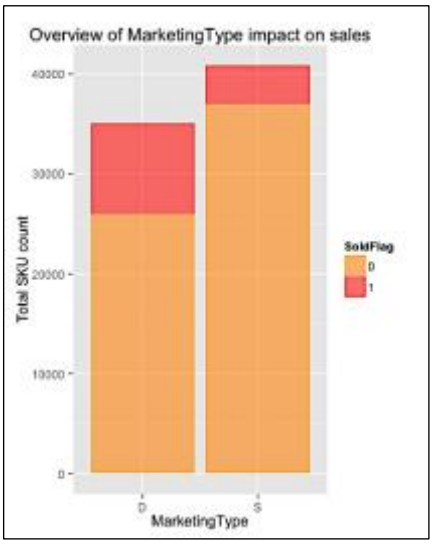

Figura 5: Impacto de *MarketingType* sobre as vendas

A figura 5 mostra uma representação gráfica da distribuição das vendas sobre cada tipo de *marketing*. É possível observar que existem mais produtos comercializados com o tipo de *marketing* “S”. Contudo, é ainda possível observar que o tipo de *marketing* “D” é o que possui um menor numero de produtos não vendidos e um maior numero de produtos vendidos. De facto, o tipo de *marketing* “D” tem mais do dobro de produtos vendidos do que o tipo de marketing “S”.

5.2.2 Impacto de *NewReleaseFlag* sobre *SoldFlag*

A variável *NewReleaseFlag* indica se um produto teve anteriormente um novo lançamento (ou seja, mais que um lançamento). A análise do impacto que novos lançamentos apresentam sobre as vendas permite obter a sua relevância.

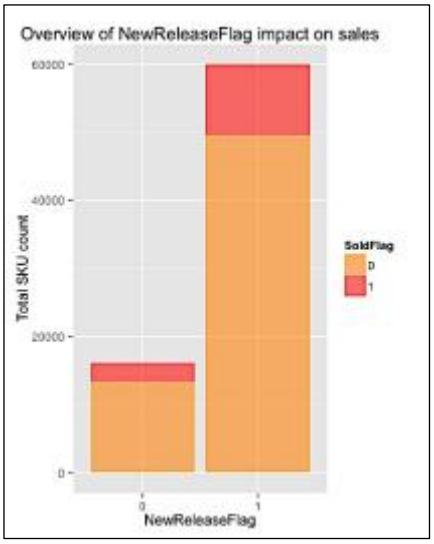

Figura 6: Impacto de NewReleaseFlag sobre as vendas

A figura 5 mostra que existem aproximadamente três vezes mais produtos que tiveram um novo lançamento em relação aos que não tiveram. Comparando o numero de vendas de

cada caso, os produtos que tiveram um novo lançamento possuem cerca de quatro vezes mais produtos vendidos em comparação aos produtos que não tiveram um novo lançamento.

5.2.3 Impacto de *StrengthFactor* sobre as vendas

StrengthFactor representa a resistência dos produtos. Verificar para que valores de resistência é que ocorrem mais vendas torna-se essencial para determinar se esta variável é relevante para pertencer aos modelos de aprendizagem. A figura 7 apresenta duas representações do impacto desta variável sobre as vendas.

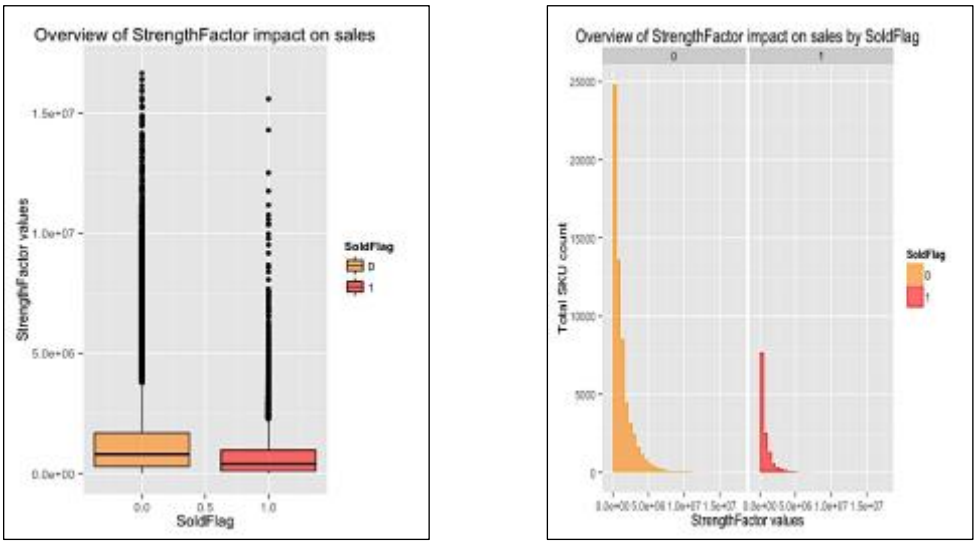

Figura 7: Impacto de StrengthFactor sobre as vendas

Da figura acima retira-se que à medida que o fator de resistência aumenta, o numero de vendas diminui, ou seja, os produtos com menor fator de resistência são mais vendidos do que aqueles com grande fator de resistência.

5.2.4 Impacto de *ReleaseYear* sobre as vendas

O ano de lançamento de um produto pode ser um fator muito relevante para a venda desse produto. Na figura 8 é possível observar a influencia que esta variável tem sobre a venda de produtos.

À medida que os anos aumentam o numero de vendas também aumenta até cerca do ano 2013, tendendo depois a diminuir para anos superiores.

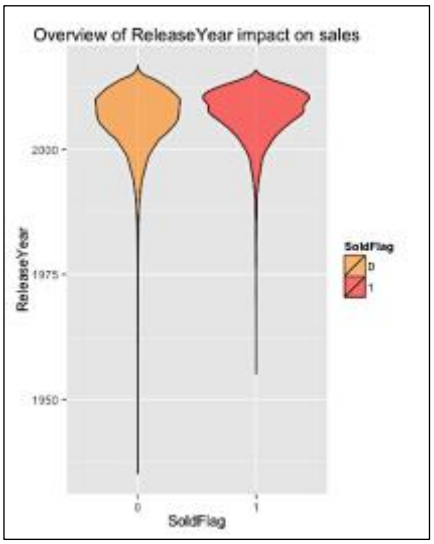

Figura 8: Impacto de ReleaseYear sobre as vendas

5.2.5 Impacto das variáveis de preços sobre as vendas :: *PriceReg*, *LowUserPrice* e *LowNetPrice*

O preço de um produto pode influenciar significativamente a venda desse produto. Dado que o conjunto de dados possui três variáveis representativas dos preços foi efetuada uma análise inicial do impacto que cada uma das variáveis possui sobre as vendas. As figuras 9, 10 e 11 mostram esse impacto.

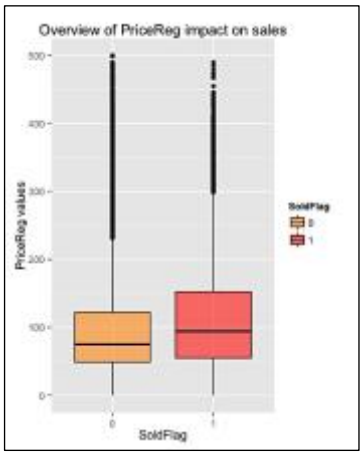

Figura 9: Impacto de PriceReg sobre as vendas

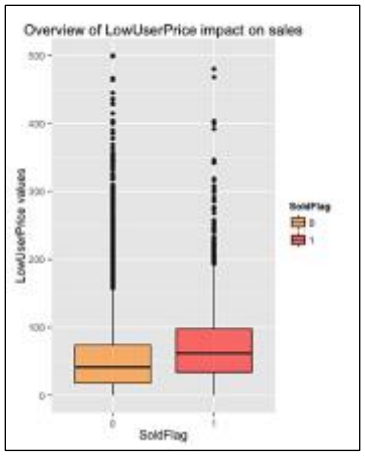

Figura 10: Impacto de LowUserPrice sobre as vendas

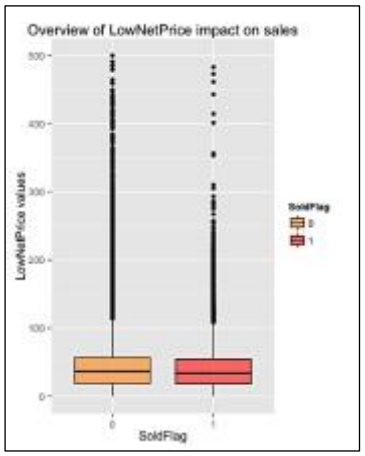

Figura 11: Impacto de LowNetPrice sobre as vendas

Atendendo que as figuras acima possuem a mesma escala, a variável *PriceReg* é a que apresenta um maior numero de vendas, sem contar com as vendas consideradas *outliers*, quando comparada com a quantidade de produtos não vendidos. As variáveis *LowUserPrice* e *LowNetPrice* parecem ser semelhantes na quantidade de vendas e não vendas.

Com esta representação da influencia de cada variável de preço sobre as vendas pode-se concluir que a variável *PriceReg* é possivelmente a melhor variável a ser utilizada para representar o preço dos produtos. Contudo, as restantes variáveis de preços, aparentemente semelhantes na quantidade de vendas e não vendas, podem ser significativas.

Por forma a compreender melhor a distinção entre as variáveis de preços e o seu impacto sobre as vendas, vai ser efetuada uma análise mais detalhada a cada uma no próximo capítulo.

5.3. Análise detalhada das variáveis de preços :: *PriceReg*, *LowUserPrice* e *LowNetPrice*

Vai ser agora efetuada uma análise mais detalhada das variáveis *PriceReg*, *LowUserPrice* e *LowNetPrice*, que representam os vários tipos de preços dos produtos. Esta análise é efetuada sobre o conjunto de dados históricos do *dataset*.

Os resultados das funções "*pairs"* e "*cor"*, anteriormente apresentados nas figuras 2 e 3, mostram que não existe uma correlação significativa entre as três variáveis de preços nem com a venda de produtos. Contudo, como o *dataset* possui um numero considerável de entradas, a correlação que exista com as vendas, embora possa ser pequena, pode ser significativa para pertencer aos modelos.

A figura 12 apresenta a distribuição dos valores das variáveis de preços.

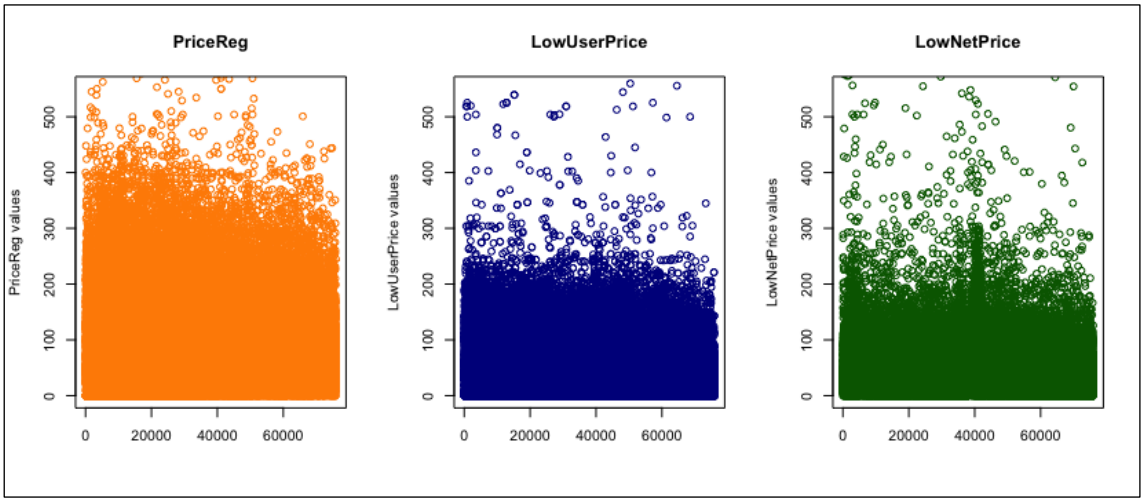

Figura 12: Distribuição dos valores das variáveis *PriceReg*, *LowUserPrice* e *LowNetPrice*

Pela figura 12 é possível observar a distribuição dos valores dos preços das variáveis de forma individual.

A figura 13 permite comparar de forma mais rigorosa os valores destas variáveis, permitindo perceber qual a variável que possui o preço mais e menos elevado.

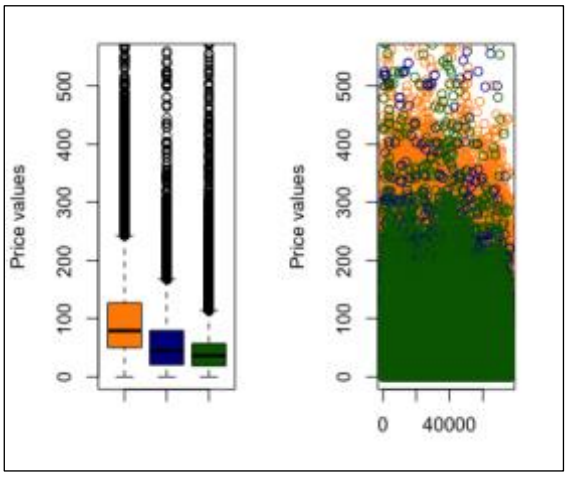

Figura 13: Comparação da distribuição dos valores das variáveis *PriceReg*, *LowUserPrice* e *LowNetPrice*
Laranja – *PriceReg;* Azul – *LowUserPrice*; Verde - *LowNetPrice*

A figura 13 contém duas representações da distribuição dos valores dos vários tipos de preços. Do lado esquerdo é apresentado um *boxplot* desses valores de cada variável. Do lado direito foi efetuada a junção dos três gráficos da figura 12 num único.

As figuras 12 e 13 mostram assim que a variável *PriceReg* é a que possui valores mais elevados dos preços, seguida pela variável *LowUserPrice* e por fim a variável *LowNetPrice* que é a que possui valores menores dos preços.

Apesar de a análise até agora efetuada permitir inferir que possivelmente a variável *PriceReg* será a melhor a ser utilizada para representar o preço, torna-se necessário efetuar uma análise mais detalhada à correlação de cada variável de preço sobre as vendas de produtos. Para tal foi efetuado um *t-test* sobre cada uma das variáveis em função das vendas.

5.3.1 *T-test* sobre as variáveis de preços

Ao efetuar este tipo de análise sobre as variáveis representativas dos preços leva a que se consiga obter com maior precisão a correlação destas variáveis com a venda dos produtos. Os resultados das funções “pairs” e “cor” anteriormente apresentados mostram que não existe uma correlação significativa entre estas variáveis e as vendas, contudo, a que existe pode já ser relevante dado o tamanho do *dataset*.

Para efetuar esta análise foi necessário começar por definir a hipótese nula.

- **Hipótese Nula:**
  A venda de produtos não está correlacionada, de forma significativa, com a variável de preço utilizada.

Consoante os valores obtidos de *p-value,* caso seja 0 então a hipótese nula deve ser aceite, não existindo uma correlação significativa entre as variáveis. Caso seja inferior a 0.05 (ou

seja, 5% do nível de significância) então a hipótese nula é descartada e conclui-se que existe uma correlação significativa entre as variáveis. Caso seja superior a 0.05 então a hipótese nula não pode ser descartada uma vez que não existem evidências suficientes que sustentem que existe uma relação linear significativa entre as variáveis.

A tabela a baixo mostra o resultado desta análise para cada variável de preço.

PriceReg	***LowUserPrice***	***LowNetPrice***
Welch Two Sample t-test data: PriceReg.Sale and PriceReg.NoSale t = 21.8525, df = 17675.87, **p-value < 2.2e-16** alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: 16.02577 19.18397 sample estimates: mean of x mean of y 113.3249 95.7200	Welch Two Sample t-test data: LowUserPrice.Sale and LowUserPrice.NoSale t = 22.554, df = 30980.47, **p-value < 2.2e-16** alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: 14.86834 17.69854 sample estimates: mean of x mean of y 70.20726 53.92382	Welch Two Sample t-test data: LowNetPrice.Sale and LowNetPrice.NoSale t = -6.2674, df = 48539.91, **p-value = 3.702e-10** alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -6.216039 -3.254352 sample estimates: mean of x mean of y 43.26073 47.99592

Tabela 11: Resultado de t-test sobre as variáveis de preços em função das vendas

Atendendo aos resultados obtidos dos valores de *p-value* das variáveis, todas possuem uma correlação significativa com as vendas dos produtos, ao contrário do que é apresentado pelas funções “pairs” e “cor”.

Uma vez que todas as variáveis de preços são significativas, torna-se necessário decidir qual ou quais utilizar nos modelos a desenvolver. Neste sentido, a variável *LowNetPrice* é descartada por possuir o maior valor de *p-value* do *t-test*. As variáveis *PriceReg* e *LowUserPrice* possuem o mesmo valor de *p-value* logo são igualmente significativas para as vendas. Por forma a escolher qual variável a utilizar de entre as duas, considerou-se a análise previamente efetuada a cada uma, onde a variável *PriceReg* é a que possui valores mais elevados de preços. Logo, esta variável é a escolhida para representar o preço dos produtos uma vez que é mais ou tão significativa do que as restantes variáveis, como foi visto pelo resultado do *t-test*, e possui maiores valores de preços o que proporciona, potencialmente, mais lucro na venda dos produtos.

Note-se que toda esta análise efetuada em redor das variáveis de preços deve-se em grande parte por não haver informação que distinga estas variáveis.

6. Modelos

Vão ser efetuados agora os modelos de aprendizagem automática para cada uma das questões de análise anteriormente definidas.

Todos os processos de compreensão e preparação dos dados efetuados anteriormente têm agora o seu valor representado na medida em que é através desta análise aos dados previamente efetua que o processo de modelagem se torna mais simples na medida em que já são conhecidas as descrições das variáveis, e quais as de maior interesse para análise encontrando-se já preparadas para poderem ser utilizadas nos modelos.

6.1. Técnicas de Modelação

Cada uma das questões de análise devem ser representadas por um modelo de aprendizagem automática que faz uso de técnicas específicas do tipo de modelo escolhido. No sentido de escolher o tipo de modelo a ser utilizado que mais precisamente caracterize as perguntas de análise, foram definidos os modelos para cada questão:

<table>
<tr><th>Pergunta de Análise</th><th>Modelo Escolhido</th><th>Tipos do Modelo Desenvolvidos</th><th>Outras Considerações</th></tr>
<tr><td>1</td><td>-</td><td>-</td><td>Respondida pela análise efetuada ás variáveis nas fases de compreensão e preparação de dados.</td></tr>
<tr><td>2</td><td>Classificação</td><td>Regressão Logística, Árvores de Decisão, Random Forests</td><td rowspan="4">Treino sobre uma amostragem dos dados históricos e teste sobre os restantes dados históricos.</td></tr>
<tr><td>3</td><td rowspan="3">Regressão</td><td rowspan="3">Regressão Linear, Árvores de Decisão, Random Forests</td></tr>
<tr><td>4</td></tr>
<tr><td>5</td></tr>
</table>

Tabela 12: Modelos de aprendizagem automática desenvolvidos para dar resposta a cada pergunta de análise

Os modelos escolhidos para cada questão de análise vão de acordo com o objetivo final da questão, quer se queira obter uma probabilidade ou a previsão de um valor numérico.

Para além dos modelos desenvolvidos para cada questão, foi ponderada a utilização de outros modelos, como clustering e regras de associação. Contudo, dada a natureza das questões de análise e sobretudo a natureza e descrição dos dados do *dataset*, os mesmos não são propícios para a aplicação deste tipo de modelos.

6.2. Definição das Condições de Teste

Para cada modelo desenvolvido é necessário definir a forma como vai ser avaliada a precisão do modelo e o conjunto de dados utilizados para esse efeito.

Neste sentido, todos os modelos são treinados com 70% dos dados históricos do *dataset* e é testado o seu nível de precisão com os restantes 30% dos dados históricos. Como os modelos são de classificação e regressão (modelos supervisionados), a forma como os mesmos são avaliados são:

- **Modelos de classificação:**
  Matriz de confusão, precisão (*accuracy*), sensibilidade (*sensitivity*), especificidade (*specificity*) e curva ROC.

- **Modelos de regressão:**
  Erros MAE, RMSE, Correlação das variáveis e R-Quadrado.

Sempre que o primeiro resultado obtido pela aplicação dos modelos não seja o melhor, o modelo é repetido, com ajustes ao mesmo, até que os valores que o classificam sejam aceitáveis, ou até que não exista uma mudança significativa na capacidade de previsão do modelo.

6.3. Construção e Avaliação dos Modelos

Vão ser agora apresentados os modelos desenvolvidos para cada questão de análise, apresentando a forma como foram implementados e avaliando a sua capacidade de previsão.

Relembra-se que a questão de análise 1 foi já analisada nos processos de compreensão e preparação dos dados.

6.3.1 Modelo questão de análise 2

Por forma a relembrar o que se pretende obter com a implementação deste modelo, relembra-se a questão de análise que lhe deu origem.

"Probabilidade de venda de cada produto (consoante o conjunto de variáveis escolhido e o melhor modelo obtido)."

O modelo desenvolvido utiliza um conjunto de variáveis do *dataset,* que foram consideradas as mais relevantes para pertencerem ao modelo de aprendizagem, de acordo com a análise dos dados previamente efetuada. As variáveis utilizadas são assim:

- *SoldFlag* (variável a ser prevista)
- *MarketingType*
- *NewReleaseFlag*
- *StrengthFactor*

- *ReleaseYear*
- *PriceReg*
- *ItemCount*

Para obter a probabilidade de venda de um produto utilizou-se vários tipos de modelos de classificação (apresentados na tabela 12) por forma a comparar os resultados de cada um e assim decidir sobre qual o melhor modelo a ser aplicado.

6.3.1.1 Modelo de Regressão Logística

O modelo de regressão logística foi aplicado recorrendo à função abaixo que faz uso de uma família apropriada para modelos de classificação e dos dados de treino:

```
logistic_model <- glm(SoldFlag ~ MarketingType + NewReleaseFlag + StrengthFactor + ReleaseYear + PriceReg + ItemCount,
family = binomial(link = 'logit'), data = train_sales)
```

Foi depois utilizado o resultado da construção do modelo acima para efetuar a previsão da probabilidade de venda sobre os produtos do conjunto de dados de teste para que seja possível avaliar o modelo desenvolvido.

```
predict(logistic_model, test_sales, type = "response")
```

Nos resultados obtidos, todos aqueles que sejam superiores ou iguais a 0.5 são classificados como 1 (venda) e aqueles que são inferiores a 0.5 são classificados como 0 (não venda). Na tabela 13 é apresentada a matriz de confusão e algumas métricas relevantes para a análise dos resultados obtidos.

Matriz de confusão	Precisão	Sensibilidade	Especificidade
Reference Prediction 0 1 0 18708 3624 1 186 280	0.8329	0.99016	0.07172

Tabela 13: Matriz de confusão e métricas obtidas

Como é possível observar pela tabela 13, os valores de precisão e sensibilidade são ótimos teoricamente. Contudo, a especificidade é extremamente baixa devido ao não balanceamento da variável *SoldFlag*, que foi já analisado na compreensão dos atributos.

Para colmatar este não balanceamento da variável *SoldFlag* pode ser efetuada uma reamostragem do *dataset*, recorrendo a vários métodos especializados para esse efeito. Faz assim sentido que seja efetuada a reamostragem dos dados de treino recorrendo a vários métodos e comparar os resultados de previsão entre eles e o modelo que utiliza o *dataset* original. As técnicas de reamostragem dos dados escolhidas foram:

- Oversampling

- Undersampling
- Both oversampling and undersampling
- SMOTE

Foram então desenvolvidos novos modelos de regressão logística utilizando os dados de treino reamostrados com cada técnica acima referida. Foi depois aplicado os novos modelos, agora balanceados na variável *SoldFlag*, para efetuar a previsão da probabilidade de venda dos produtos. No anexo 2 é possível observar a tabela 35 que apresenta a matriz de confusão e os resultados das métricas escolhidas para avaliação dos novos modelos desenvolvidos.

A tabela 14 mostra estas métricas e as métricas do modelo originalmente efetuado. A figura 14 apresenta uma representação gráfica da curva ROC de cada modelo.

Modelo	Precisão	Sensibilidade	Especificidade	ROC
Original	0.8329	0.9901	0.0717	0.745
Oversampling	0.6998	0.7078	0.6611	0.746
Undersampling	0.7007	0.7086	0.6624	0.746
Both Over. & Under.	0.6995	0.7072	0.6624	0.747
SMOTE	0.7448	0.7807	0.5712	0.743

Tabela 14: Valores das métricas obtidas para cada modelo de regressão logística

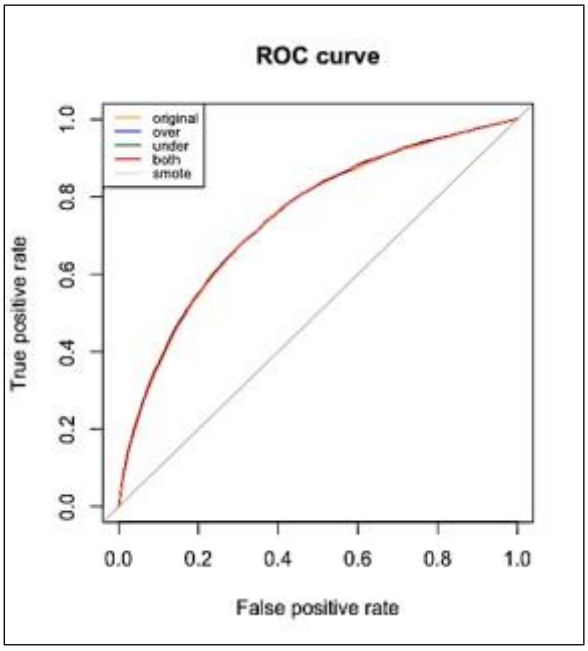

Figura 14: Representação ROC dos modelos de regressão logística

A partir da tabela e figura 14 é possível concluir que os modelos são extremamente semelhantes à exceção do modelo original que possui baixa especificidade. Desta forma, a escolha de um modelo está dependente do que se pretende maioritariamente, ou melhor precisão ou melhor sensibilidade e especificidade.

Para efeitos de comparação deste modelo com outros a serem desenvolvidos, vai ser escolhido o modelo que utiliza o método de reamostragem do *dataset* SMOTE.

6.3.1.2 Modelo de Árvores de Decisão

A construção deste modelo foi efetuada recorrendo à função abaixo que utiliza um método apropriado para problemas de classificação:

```
tree_model <- rpart(SoldFlag ~ MarketingType + NewReleaseFlag + StrengthFactor + ReleaseYear + PriceReg + ItemCount, data = train_sales, method = "class")
```

De todas as variáveis utilizadas no modelo, a aplicação desta função de árvores de decisão apenas utilizou três das variáveis devido ao nível de importância que a própria dá às variáveis, como é possível constatar pela figura 15 que apresenta a representação gráfica da árvore formada.

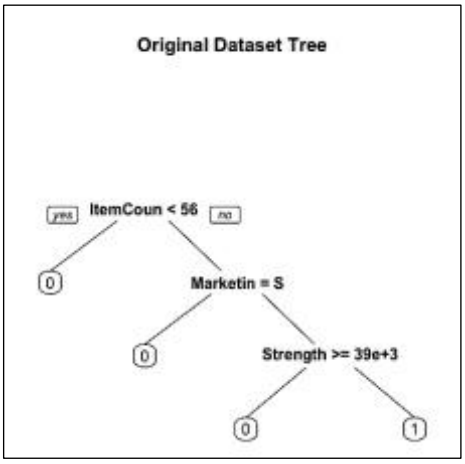

Figura 15: Árvore de decisão obtida

Foi depois utilizado o resultado da construção do modelo acima para efetuar a previsão da probabilidade de venda sobre os produtos do conjunto de dados de teste para que seja possível avaliar o modelo desenvolvido. A tabela 15 mostra a matriz de confusão e as métricas relevantes para avaliação da capacidade de previsão do modelo.

Matriz de confusão	Precisão	Sensibilidade	Especificidade
Reference Prediction 0 1 0 18728 3639 1 166 265	0.8331	0.99121	0.06788

Tabela 15: Matriz de confusão e valores obtidos das métricas

Tal como no modelo de regressão logística, este também apresenta um valor de especificidade muito baixo devido ao não balanceamento da classe *SoldFlag*. Para colmatar este problema, foram criados novos modelos que utilizam dados reamostrados a partir das mesmas técnicas anteriormente utilizadas.

A figura 16 apresenta a representação visual das novas árvores obtidas para cada um dos novos modelos desenvolvidos.

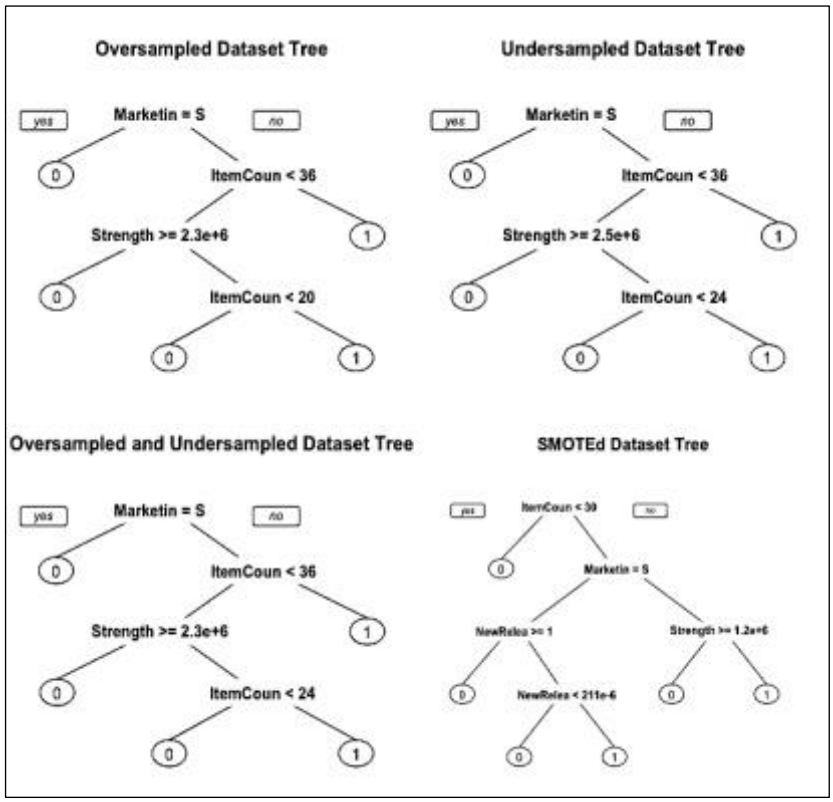

Figura 16: Árvores de decisão de cada modelo com reamostragem dos dados

As novas árvores utilizam diferentes variáveis entre elas e a árvore do modelo original devido ao fator de importância que cada uma atribui às variáveis. Foram aplicados os novos modelos para efetuar a previsão da probabilidade de venda dos produtos cujos resultados das métricas são apresentados na tabela 16 em conjunto com os valores das curvas ROC. A figura 17 apresenta uma representação gráfica da curva ROC de cada modelo.

Modelo	Precisão	Sensibilidade	Especificidade	ROC
Original	0.8331	0.99121	0.06788	0.530
Oversampling	0.7065	0.7222	0.6304	0.676
Undersampling	0.7200	0.7422	0.6127	0.677
Both Over. & Under.	0.7225	0.7460	0.6089	0.677
SMOTE	0.7785	0.8372	0.4946	0.666

Tabela 16: Valores das métricas para cada modelo de árvore de decisão

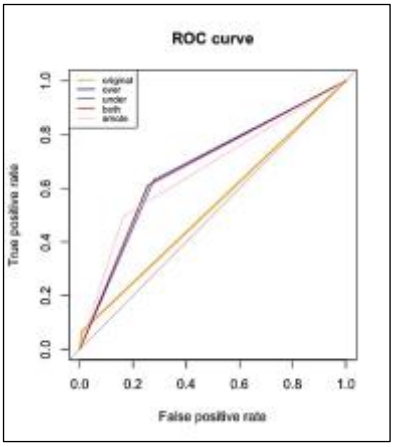

Figura 17: Curvas ROC dos modelos de árvores de decisão

Os modelos com os dados reamostrados possuem um valor de especificidade aceitável em comparação com o modelo original que também é refletido nas curvas ROC da figura 17 que

possuem um aumento significativo em relação ao modelo original. Contudo, os valores das métricas entre os vários modelos com reamostragem dos dados são praticamente idênticos.

Para efeitos de comparação deste modelo com outros a serem desenvolvidos, vai ser escolhido o modelo que utiliza o método de reamostragem do *dataset* SMOTE.

6.3.1.3 Modelo de Random Forests

O ultimo modelo desenvolvido para dar resposta a esta questão de análise é um modelo baseado em random forests. Este modelo foi construído recorrendo à função abaixo, com a geração de 1000 árvores. No anexo 2, figura 20 apresenta uma representação gráfica dos erros deste modelo.

```
randomF_model <- randomForest(SoldFlag ~ MarketingType + NewReleaseFlag + StrengthFactor + ReleaseYear + PriceReg + ItemCount, data = train_sales, ntree=1000)
```

Foi depois utilizado o resultado da construção do modelo acima para efetuar a previsão da probabilidade de venda sobre os produtos do conjunto de dados de teste para que seja possível avaliar o modelo desenvolvido. A tabela 17 mostra a matriz de confusão e as métricas relevantes para avaliação da capacidade de previsão do modelo.

Matriz de confusão	Precisão	Sensibilidade	Especificidade
Reference Prediction 0 1 0 18605 3467 1 289 437	0.8352	0.9847	0.1119

Tabela 17: Matriz de confusão e valores obtidos das métricas

Como seria de esperar, tal como nos modelos anteriores, o valor de especificidade é muito baixo. Foram, portanto, desenvolvidos novos modelos com dados reamostrados utilizando as mesmas técnicas dos modelos anteriores. O anexo 2 figura 21 apresenta uma representação gráfica dos erros dos novos modelos.

Os resultados das métricas obtidas com os novos modelos são apresentados na tabela 18 em conjunto com os valores das curvas ROC. A figura 18 apresenta uma representação gráfica das curvas ROC de cada modelo.

Modelo	Precisão	Sensibilidade	Especificidade	ROC
Original	0.8352	0.9847	0.1119	0.548
Oversampling	0.7653	0.8102	0.5484	0.679
Undersampling	0.6921	0.6936	0.6852	0.689
Both Over. & Under.	0.7389	0.7677	0.5999	0.684
SMOTE	0.794	0.8717	0.4180	0.645

Tabela 18: Valores das métricas para cada modelo de random forests

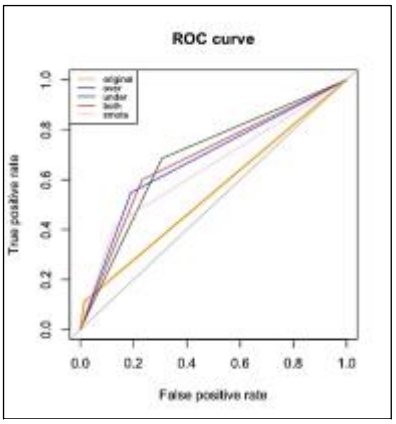

Figura 18: Curvas ROC dos modelos de random fosrests

Os valores da especificidade dos modelos com os dados reamostrados são consideravelmente melhores do que o modelo original o que é refletido na curva ROC como é possível observar pela figura 18. Os modelos com reamostragem dos dados são, no entanto, muito semelhantes entre eles. Para efeitos de comparação deste modelo com outros já desenvolvidos, vai ser escolhido o modelo que utiliza o método de reamostragem do *dataset* SMOTE.

6.3.1.4 Escolha do Melhor Modelo

Foram desenvolvidos três modelos de aprendizagem automática por forma a decidir qual o mais adequado a ser utilizado na questão de análise. Em cada modelo desenvolvido foram efetuadas várias variações por forma a melhorar os resultados obtidos, tendo sido escolhida a melhor otimização de cada modelo. Falta agora comparar os melhores resultados obtidos em cada modelo por forma a decidir sobre qual utilizar.

Como forma de comparação recorreu-se à análise da curva ROC dos melhores modelos considerados, que pode ser observado na figura 19.

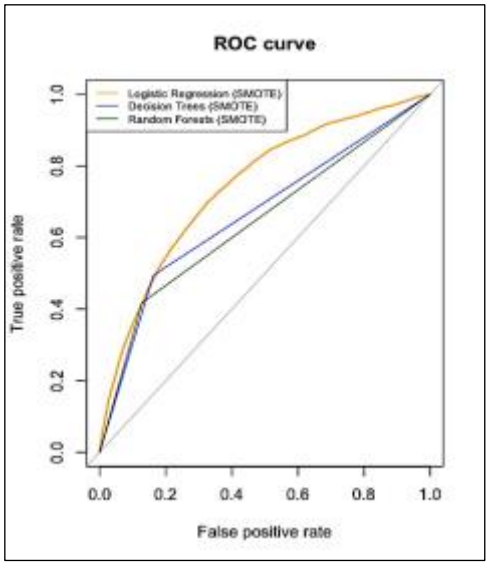

Figura 19: Curvas ROC dos melhores modelos conseguidos.

Os melhores modelos conseguidos foram, como referido ao longo da análise, aqueles que utilizam o método SMOTE para efetuar a reamostragem do *dataset*, obtendo assim um conjunto de dados balanceado. Pela figura 19 é possível observar que o modelo de regressão logística é o que apresenta a maior "*area under the curve"*, logo, é este modelo que é escolhido para ser utilizado na previsão da probabilidade de venda de cada produto.

Pode então ser facilmente aplicado o modelo de regressão logística com reamostragem SMOTE sobre os dados ativos do *dataset* para prever a probabilidade de venda de cada produto neles contido.

6.3.2 Modelo questão de análise 3

Por forma a relembrar o que se pretende obter com a implementação deste modelo, a questão de análise é apresentada abaixo.

"Quantidade esperada de venda de cada produto (de acordo com o melhor modelo conseguido)."

Para dar resposta a esta questão de análise foram ponderadas duas abordagens distintas:

1. Utilizar um modelo de classificação sendo necessário transformar os valores da variável *SoldCount* em categorias que seriam representativas de intervalos de valores de venda.
2. Utilizar um modelo de regressão linear para obter um valor numérico previsto da quantidade a ser vendida de cada produto.

Utilizando a opção 2 acima referida, foi necessário começar por decidir qual o conjunto de variáveis relevantes a pertencerem ao modelo de aprendizagem, de acordo com a análise dos dados previamente efetuada. As variáveis utilizadas são assim:

- *SoldCount* (variável a ser prevista)
- *MarketingType*
- *NewReleaseFlag*
- *StrengthFactor*
- *ReleaseYear*
- *PriceReg*
- *ItemCount*

Para obter a probabilidade de venda de um produto utilizou-se vários tipos de modelos de regressão (apresentados na tabela 12) por forma a comparar os resultados de cada um e assim decidir sobre qual o melhor modelo a ser aplicado.

6.3.2.1 Modelo de Regressão Linear

O modelo de regressão linear foi desenvolvido recorrendo à função abaixo, que utiliza o conjunto de variáveis indicadas e os dados de treino (amostragem de 70% dos dados históricos).

```
linear_r_model <- lm(SoldCount ~ MarketingType + NewReleaseFlag + StrengthFactor + ReleaseYear + PriceReg + ItemCount,
data = train_sales)
```

As figuras 20 e 21 apresentam a visualização de um conjunto de métricas características do tipo de modelo desenvolvido que ajudam a perceber a qualidade do modelo e a forma como se adequa aos dados.

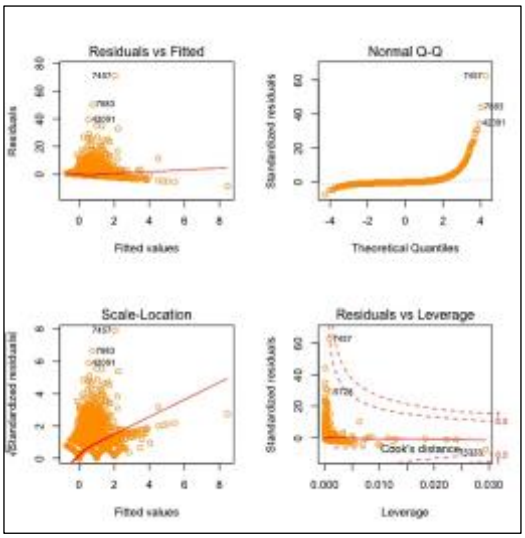

Figura 21: Métricas características do modelo de regressão linear

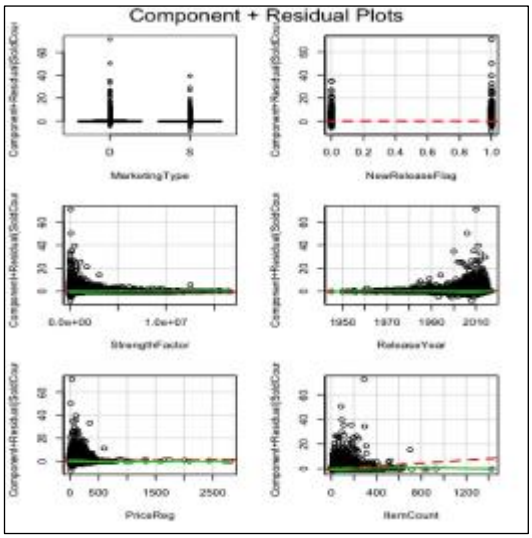

Figura 21: Ajuste dos erros em relação ao modelo linear de cada variável

A figura 20 mostra que os valores “residuals vs fitted” e “residuals vs leverage” apresentam um comportamento aceitável, contudo, os valores de “normal q-q” e “scale-location” mostram a presença de não linearidade, muito provavelmente devido à presença de uma grande quantidade de *outliers* nos dados (cuja análise foi efetuada anteriormente). A figura 21 mostra o ajuste dos erros de cada variável em redor do modelo linear. Com a exceção da variável *ItemCount* que apresenta um pouco de não linearidade, todas as variáveis possuem um bom ajuste dos erros.

Foi utilizado este modelo para efetuar a previsão da quantidade de venda de cada produto do conjunto de dados de teste para que seja possível avaliar o modelo. No anexo 3, figura 25 apresenta a comparação dos valores “*actual vs predicted*” sobre os dados de treino e de teste. A tabela 19 mostra os valores dos erros e métricas consideradas relevantes para avaliar o modelo.

MAE	RMSE	CORRELATION	R2
0.38218265	1.11959240	0.24076474	0.05796766

Tabela 19: Erros MAE e RMSE e métricas Corr. e r-squared do modelo

Os resultados acima apresentados mostram que os erros MAE e RMSE são aceitáveis tendo em conta que o que se pretende determinar é uma quantidade de venda. Por outro lado, o valor de *r-squared* é extremamente baixo.

Por forma a tentar melhorar o valor de *r-squared* podem ser eliminados os valores mais extremos que podem ter um impacto significativo na performance do modelo. Neste sentido foi utilizada a distância de Cook para encontrar os valores extremos utilizando a função abaixo. Os extremos podem ser visualizados nas figuras 22 e 23.

```
cutoff <- 4/((nrow(train_sales) - length(linear_r_model$coefficients) - 2))
```

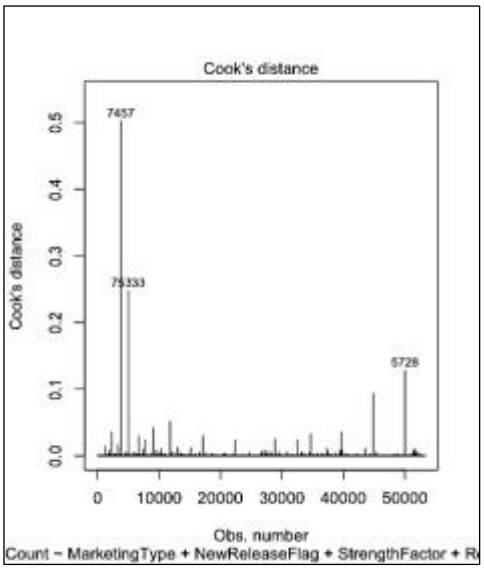

Figura 23: Valores extremos da distância de Cook

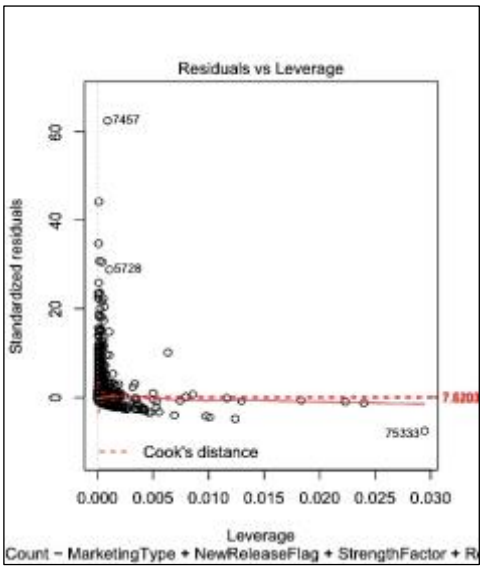

Figura 23: Valores extremos da distância de Cook

Os valores extremos são retirados dos dados de treino e o modelo é recalculado. Este processo foi efetuado um numero de vezes até que fossem obtidos os melhores valores possíveis. Durante este processo, a variável *NewReleaseFlag* foi descartada do modelo uma vez que deixou de ser significativa como pode ser observado pelo extrato abaixo da aplicação de *summary()* sobre o modelo.

```
Coefficients:
                    Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)       -1.592e+01  1.450e+00  -10.977  <2e-16 ***
MarketingTypeS    -3.225e-01  8.592e-03  -37.532  <2e-16 ***
NewReleaseFlag    -2.001e-02  1.028e-02   -1.945   0.0517 .
StrengthFactor    -4.644e-08  3.020e-09  -15.379  <2e-16 ***
ReleaseYear        8.077e-03  7.229e-04   11.172  <2e-16 ***
PriceReg           4.685e-04  5.362e-05    8.738  <2e-16 ***
ItemCount          5.197e-03  1.214e-04   42.792  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A tabela 20 mostra a comparação dos valores dos erros e das métricas para o modelo original e o modelo final onde se tentou otimizar principalmente o valor de *r-squared*.

	MAE	RMSE	CORRELATION	R2
Original	0.38218265	1.11959240	0.24076474	0.05796766
Otimizado	0.37314677	1.12045399	0.23542211	0.05542357

Tabela 20: Erros MAE e RMSE e métricas Corr. e r-squared do modelo original e final

A tabela mostra que apesar de ter sido efetuado um modelo final, possivelmente otimizado pela remoção de valores extremos, o mesmo não trouxe nenhuma otimização ao modelo original. Uma possível explicação é que o numero de valores extremos removidos foram insignificantes para o numero total de valores extremos (outliers). Voltar a efetuar este modelo

com o conjunto de dados sem a presença de outliers pode resultar numa melhoria significativa do modelo. Apesar disto, os valores dos erros MAE e RMSE continuam a ser aceitáveis.

6.3.2.2 Modelo de Árvores de Decisão

A construção deste modelo foi efetuada recorrendo à função abaixo que utiliza um método apropriado para problemas de regressão:

```
tree_r_model <- rpart(SoldCount ~ MarketingType + NewReleaseFlag + StrengthFactor + ReleaseYear + PriceReg + ItemCount,
data = train_sales, method = "anova")
```

A figura abaixo mostra a representação gráfica da árvore formada por este modelo.

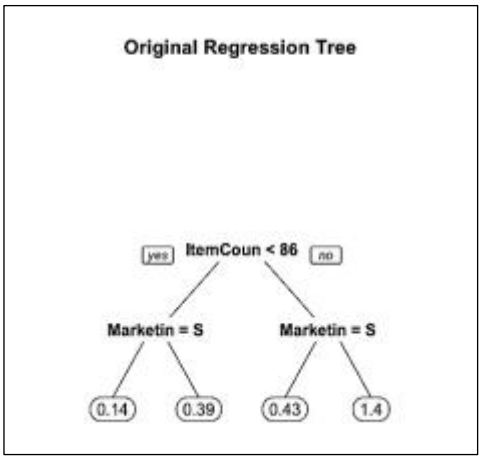

Figura 24: Árvore de decisão obtida

Foi aplicado este modelo para efetuar a previsão da quantidade esperada de venda de cada produto no conjunto de dados de teste, obtendo assim o valor dos erros e métricas relevantes para avaliação do modelo que podem ser observadas na tabela 21.

MAE	RMSE	CORRELATION	R2
0.31621195	1.13997536	0.21223295	0.04504283

Tabela 21: Erros MAE e RMSE e métricas Corr. e r-squared do modelo

Tendo em conta a tabela acima, os valores dos erros são aceitáveis. Os valores de correlação e *r-squared* são muito baixos. Observando a árvore gerada na figura 24 é ainda possível constatar que os valores gerados nas folhas não são aceitáveis uma vez que não representam os valores da quantidade de vendas de cada produto num domínio propicio para tal (as possibilidades são apenas 0 vendas e 1 venda).

O modelo foi otimizado por forma a que sejam utilizadas mais variáveis, aumentando assim o numero de folhas da árvore e consequentemente o domínio dos valores possíveis de quantidade de venda.

A nova árvore gerada por esta otimização pode ser consultada no anexo 3, figura 29.

À nova árvore gerada foi efetuada uma verificação dos erros em relação ao seu tamanho por forma a encontrar o tamanho da árvore que minimiza os erros e que mantém um domínio aceitável sobre os valores previstos.

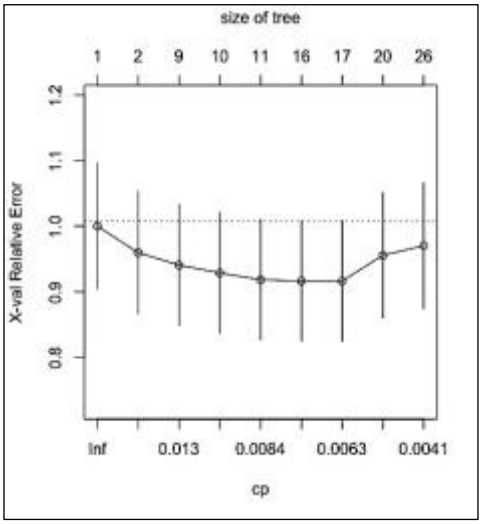

Figura 25 Erros da árvore otimizada em função do seu tamanho

A árvore otimizada foi então cortada (*pruned*) para um tamanho de 17 de acordo com a informação da imagem 25 que minimiza os erros, obtendo uma nova árvore. A árvore final encontra-se na figura a baixo.

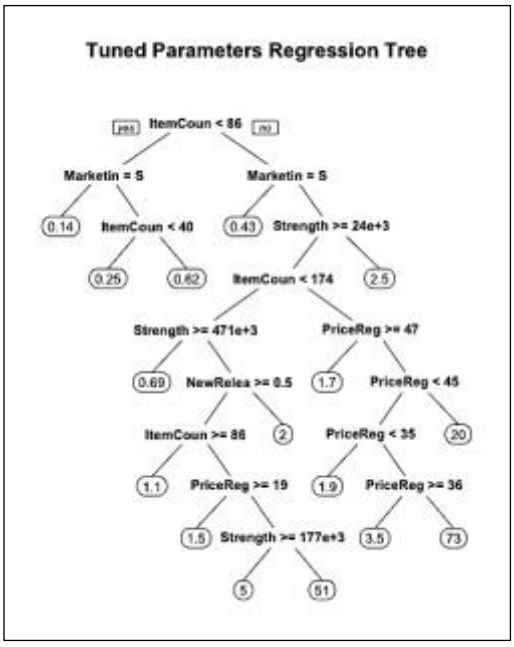

Figura 26: Árvore final do modelo

Na tabela 22 pode ser visto os valores dos erros e das métricas de avaliação do modelo em comparação com o modelo original.

	MAE	RMSE	CORRELATION	R2
Original	0.31621195	1.13997536	0.21223295	0.04504283
Otimizado	0.37805948	1.12502498	0.25273915	0.06387708

Tabela 22: Erros MAE e RMSE e métricas Corr. e r-squared do modelo original e final

Pela tabela acima, no modelo final o erro MAE aumentou e RMSE diminuiu, mas não significativamente. Os valores da correlação e *r-squared* aumentaram. Para além destas métricas, a árvore final apresenta um conjunto de valores nas folhas aceitáveis para a variável que se quer prever.

6.3.2.3 Modelo de Random Forests

O ultimo modelo desenvolvido para dar resposta a esta questão de análise é um modelo baseado em random forests. Este modelo foi construído recorrendo à função abaixo, com a geração de 500 árvores. O anexo 3, figuras 32 e 33 mostram uma representação gráfica dos erros deste modelo e da importância de cada variável, respetivamente.

```
randomF_r_model <- randomForest(SoldCount ~ MarketingType + NewReleaseFlag + StrengthFactor + ReleaseYear + PriceReg
+ ItemCount, data = train_sales)
```

A aplicação deste modelo sobre os dados de teste permitiu obter os erros e métricas de avaliação do mesmo que se encontram na tabela 23.

MAE	RMSE	CORRELATION	R2
0.3684972	1.0900140	0.3235897	0.1047103

Tabela 23: Erros MAE e RMSE e métricas Corr. e r-squared do modelo

6.3.2.4 Escolha do Melhor Modelo

Foram desenvolvidos três modelos de aprendizagem automática por forma a decidir qual o mais adequado a ser utilizado na questão de análise. Para os modelos de regressão linear e árvores de decisão foram desenvolvidas várias otimizações por forma a melhorar os resultados obtidos. Falta agora comparar os modelos por forma a decidir sobre qual utilizar. A tabela 24 mostra os resultados de cada modelo.

	MAE	RMSE	CORRELATION	R2
Regressão Linear (otimizado)	0.37314677	1.12045399	0.23542211	0.05542357
Árvores de Decisão (otimizado)	0.37805948	1.12502498	0.25273915	0.06387708
Random Forests	0.3684972	1.0900140	0.3235897	0.1047103

Tabela 24: Erros MAE e RMSE e métricas Corr. e r-squared dos vários modelos desenvolvidos

De acordo com os valores da tabela acima, o modelo baseado em random forests é o melhor modelo uma vez que possui menores valores de erros e maior correlação e *r-squared*.

Pode então ser facilmente aplicado o modelo de random forests sobre o conjunto de dados ativos do *dataset* para prever a quantidade de venda de cada produto ou de possíveis novos produtos que possam vir a fazer parte do negócio.

6.3.3 Modelo questão de análise 4

Por forma a relembrar o que se pretende obter com a implementação deste modelo, a questão de análise é apresentada abaixo.

"Quantidade apropriada de inventário para cada produto"

Pretende-se com esta questão de análise prever a quantidade de inventário de cada produto por forma a obter um valor o mais preciso possível da quantidade que cada produto deve possuir em inventário. A figura a baixo apresenta a distribuição das vendas e das quantidades em inventário dos produtos.

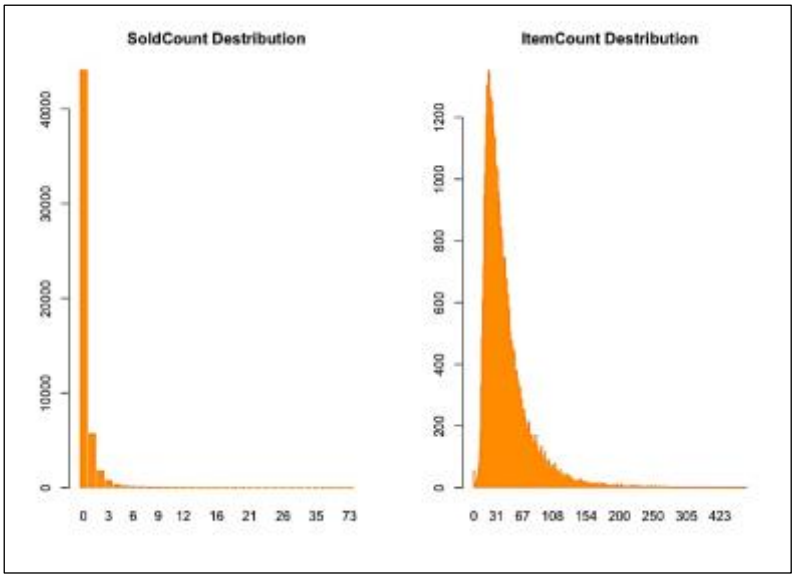

Figura 27: Distribuição das vendas e das quantidades em inventário

Como é possível observar pela figura acima e como já tinha sido constatado no capítulo de compreensão dos dados, existem muitos mais produtos que não tiveram uma venda, e aqueles que tiveram foram maioritariamente de uma e duas vendas. O valor máximo que um produto foi vendido é de 73 unidades. Observando a distribuição das quantidades em inventário existem quantidades muito altas em inventário comparando com as quantidades que são efetivamente vendidas (média da quantidade de produtos em inventário é 43).

Desta forma, a utilização dos valores em inventário para dados de treino não são os mais apropriados uma vez que os modelos vão aprender com um conjunto de dados com valores de inventário bastante mais elevados do que o numero de vendas que ocorrem. Por forma a reduzir as quantidades elevadas de produtos em inventário foi desenvolvida uma função que efetua uma reamostragem dos dados de treino, utilizando apenas aqueles que possuem valores de inventário até à soma do valor máximo de venda de um produto e a mediana dos valores de inventário.

Tendo em conta a natureza das vendas dos produtos onde existem mais produtos que não tiveram uma venda do que aqueles que tiveram, pela aplicação da função desenvolvida, consegue-se diminuir o erro de treino, de quantidades de inventário extremamente excessivas, para valores de inventário mais próximos das quantidades de venda.

A figura 28 apresenta a distribuição dos valores de inventário depois de aplicada a reamostragem dos dados. É possível ainda observar pela figura a existência de uma aproximação de uma distribuição normal.

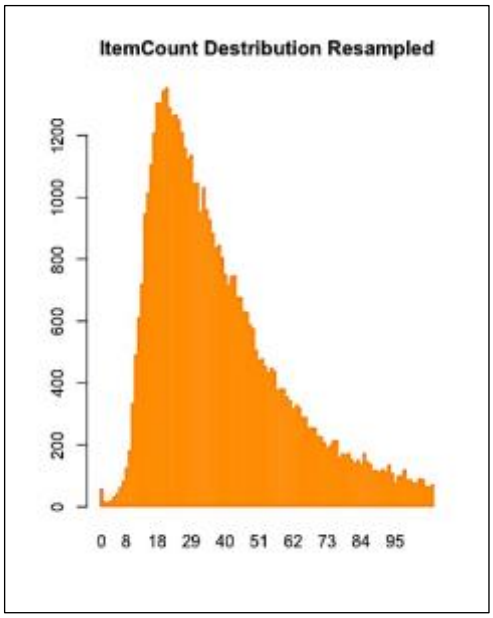

Figura 28: Distribuição das quantidades de inventário dos dados reamostrados

Foi depois necessário decidir sobre quais as variáveis relevantes a pertencerem aos modelos de aprendizagem, de acordo com a análise de dados previamente efetuada. As variáveis utilizadas são assim:

- *ItemCount* (variável a ser prevista)
- *MarketingType*
- *NewReleaseFlag*
- *StrengthFactor*
- *ReleaseYear*
- *PriceReg*

Para obter a quantidade apropriada de inventário de cada produto utilizou-se vários tipos de modelos de regressão (apresentados na tabela 12) por forma a comparar os resultados de cada um e assim decidir qual o melhor modelo a ser aplicado.

6.3.3.1 Modelo de Regressão Linear

O modelo de regressão linear foi desenvolvido recorrendo à função abaixo, que utiliza o conjunto de variáveis indicadas e os dados de treino reamostrados.

```
linear_r_model_inventario <- lm(ItemCount ~ MarketingType + NewReleaseFlag + StrengthFactor + ReleaseYear + PriceReg, data = train_sales)
```

As figuras 29 e 30 apresentam a visualização de um conjunto de métricas características do tipo de modelo desenvolvido que ajudam a perceber a qualidade do modelo e a forma como se adeque aos dados.

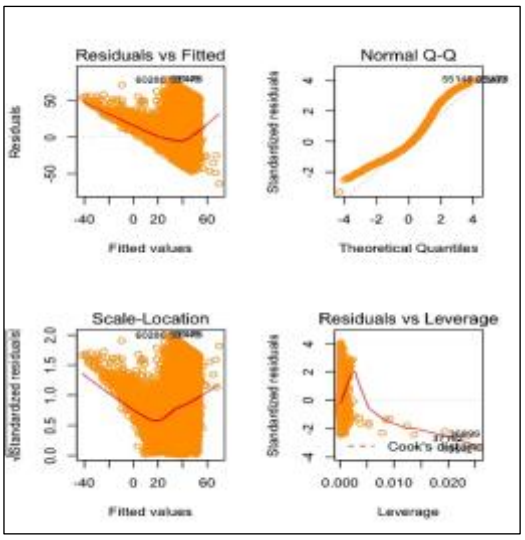

Figura 30: Métricas características do modelo de regressão linear

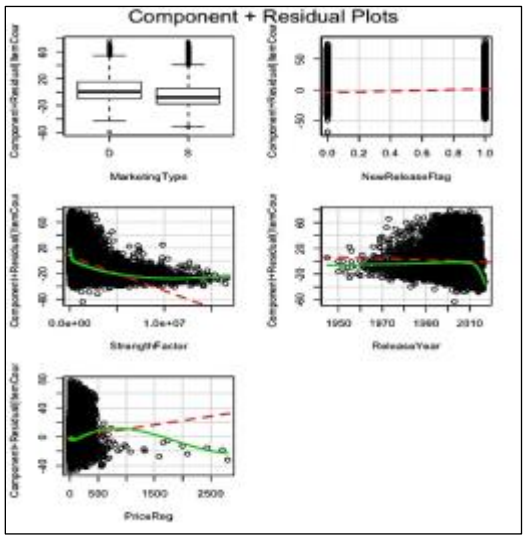

Figura 30: Ajuste dos erros em relação ao modelo linear de cada variável

A figura 29 mostra que apenas a distribuição de “normal q-q” é aceitável, mostrando a presença de não linearidade nas outras distribuições. muito provavelmente devido à presença de uma grande quantidade de *outliers* nos dados (cuja análise foi efetuada anteriormente). A figura 30 mostra o ajuste dos erros de cada variável em redor do modelo linear. Também aqui é possível observar a presença de não linearidade acentuada não existindo um bom ajuste dos erros nas variáveis.

A tabela 25 mostra os valores dos erros e das métricas relevantes para avaliação do modelo depois de aplicado aos dados de teste.

MAE	RMSE	CORRELATION	R2
19.8481007	35.7220220	0.3541459	0.1254193

Tabela 25: Erros MAE e RMSE e métricas Corr. e r-squared do modelo

Os resultados acima mostram valores de erros relativamente elevados para o contexto onde se inserem. Por outro lado, os valores de correlação e *r-squared* são dos maiores conseguidos em qualquer modelo até agora desenvolvido.

Por forma a tentar melhorar os valores dos erros, tal como na pergunta de análise anterior, podem ser eliminados os valores mais extremos que podem ter um impacto significativo na performance do modelo. Foi então utilizada a distância de Cook para encontrar os valores extremos que podem ser visualizados nas figuras 31 e 32.

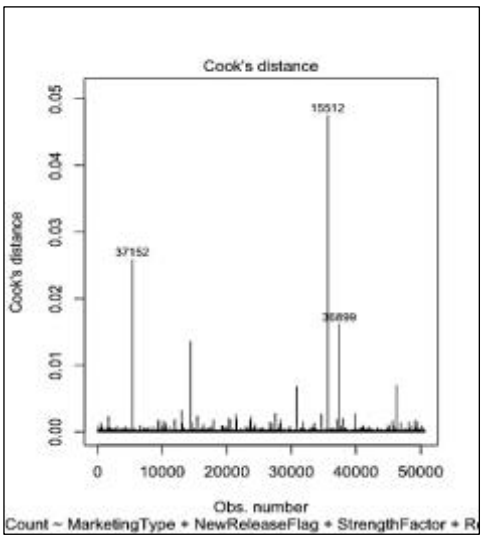

Figura 32: Valores extremos da distância de Cook

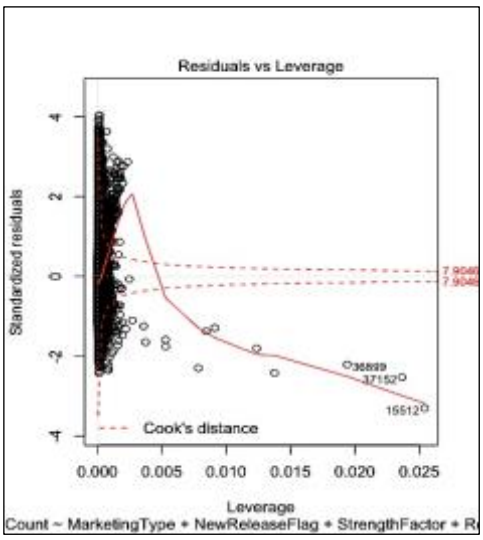

Figura 32: Valores extremos da distância de Cook

Os valores extremos são retirados dos dados de treino e o modelo é recalculado. Este processo foi efetuado um numero de vezes até que fossem obtidos os melhores valores possíveis.

A figura 33 mostra a variação dos valores de inventário previstos para os dados de teste e a figura 34 mostra a distribuição dos valores de stock originais contidos no *dataset*, dos valores previstos e da quantidade vendida de produtos. Pela figura 33 é possível ver que o intervalo de valores previstos é muito mais aceitável do que os valores que realmente existem em inventário. Isto pode ser constatado também pela figura 34 que mostra a distribuição dos valores previstos muito mais perto das quantidades vendidas.

A tabela 26 mostra a comparação dos valores dos erros e das métricas para o modelo original e o modelo final que indica que apesar de ter sido feito um modelo otimizado, o mesmo é igual ao modelo original em termos das métricas apresentadas. Contudo, estas métricas são apenas verdadeiramente relevantes se for assumido que os dados de teste possuem valores de inventário ótimos. Caso contrário, o modelo deve ser avaliado apenas no facto de que foi possível aproximar os valores previstos da quantidade de produtos vendidos.

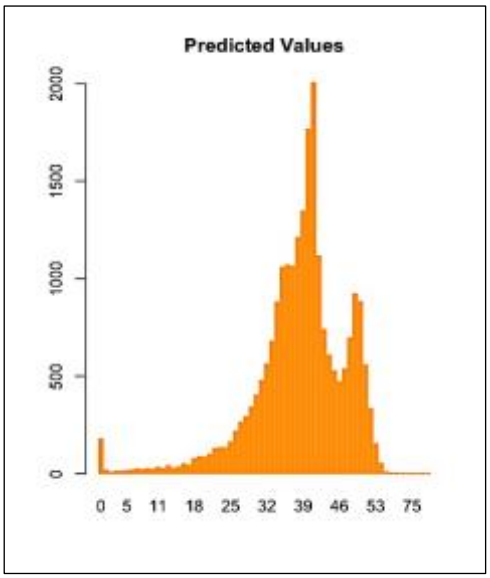

Figura 34: Valores de inventário previstos

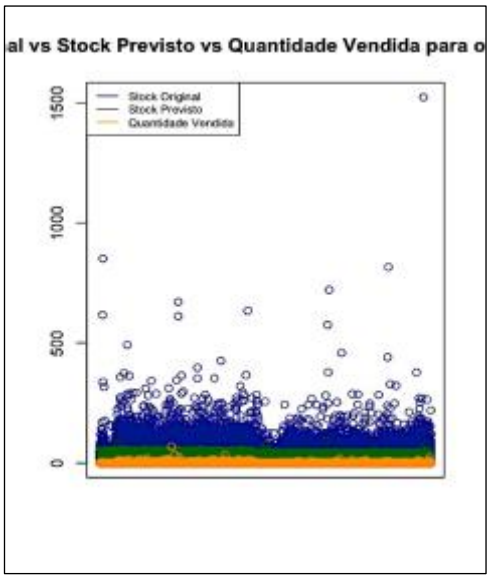

Figura 34: Stock original vs previsto vs quantidade vendida

	MAE	RMSE	CORRELATION	R2
Original	19.8481007	35.7220220	0.3541459	0.1254193
Otimizado	19.8241951	35.6924607	0.3544680	0.1256476

Tabela 26: Erros MAE e RMSE e métricas Corr. e r-squared do modelo original e final

6.3.3.2 Modelo de Árvores de Decisão

A construção deste modelo foi efetuada recorrendo à função abaixo que utiliza um método apropriado para problemas de regressão e o conjunto de dados de treino reamostrados.

```
tree_r_model_inventario <- rpart(ItemCount ~ MarketingType + NewReleaseFlag + StrengthFactor + ReleaseYear + PriceReg, data = train_sales, method = "anova")
```

A figura abaixo mostra a árvore formada por este modelo.

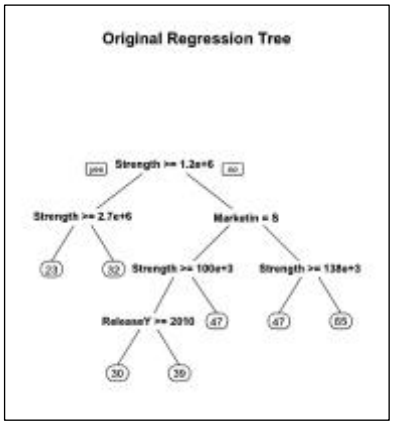

Figura 35: Árvore de decisão obtida

Pela figura 36 e tabela 27 é possível observar a distribuição e valores de erros e métricas depois de aplicar este modelo aos dados de teste.

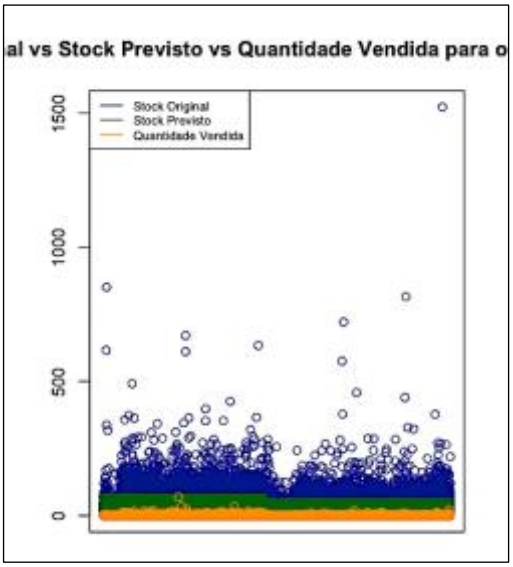

Figura 36: Stock original vs previsto vs quantidade vendida

MAE	RMSE	CORRELATION	R2
18.9238968	34.5041531	0.4384053	0.1921992

Tabela 27: Erros MAE e RMSE e métricas Corr. e r-squared do modelo

6.3.3.3 Modelo de Random Forests

O ultimo modelo desenvolvido é baseado em random forests. Foi construído pela função abaixo, com geração de 500 árvores.

```
randomF_r_model_inventario <- randomForest(ItemCount ~ MarketingType + NewReleaseFlag + StrengthFactor + ReleaseYear + PriceReg, data = train_sales)
```

Ao modelo acima criado foi efetuada uma otimização que utiliza apenas 17 árvores por forma a minimizar os erros. O anexo 4, figuras 44, 45 e 46, mostram os erros do modelo original, do modelo otimizado e a importância das variáveis, respetivamente.

A figura 37 mostra a variação dos valores de inventário previstos para os dados de teste e a figura 38 mostra a distribuição dos valores de stock originais contidos no *dataset*, dos valores previstos e da quantidade vendida de produtos.

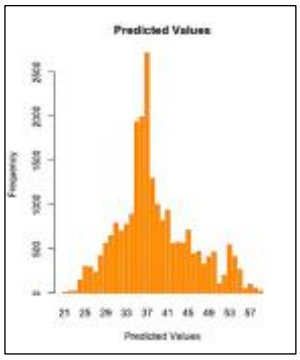

Figura 38: Valores de inventário previstos

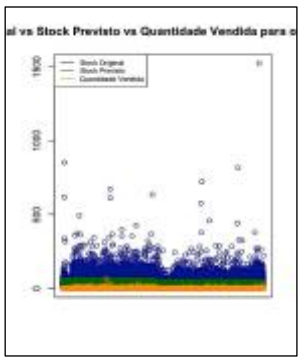

Figura 38: Stock original vs previsto vs quantidade vendida

Pela tabela 28 é possível observar a distribuição e valores de erros e métricas depois de aplicar este modelo aos dados de teste.

MAE	RMSE	CORRELATION	R2
19.1686551	35.0899393	0.4664647	0.2175893

Tabela 28: Erros MAE e RMSE e métricas Corr. e r-squared do modelo

6.3.3.4 Escolha do Melhor Modelo

A tabela a baixo mostra a comparação dos valores de erros e métricas dos vários modelos desenvolvidos.

	MAE	RMSE	CORRELATION	R2
Regressão Linear (otimizado)	19.8241951	35.6924607	0.3544680	0.1256476
Árvores de Decisão (otimizado)	18.9238968	34.5041531	0.4384053	0.1921992
Random Forests	19.1686551	35.0899393	0.4664647	0.2175893

Tabela 29: Erros MAE e RMSE e métricas Corr. e r-squared dos vários modelos desenvolvidos

Observando apenas os valores da tabela 29, o modelo de random forests é ligeiramente o melhor modelo. Deve ser ainda analisada a distribuição dos valores previstos por cada modelo.

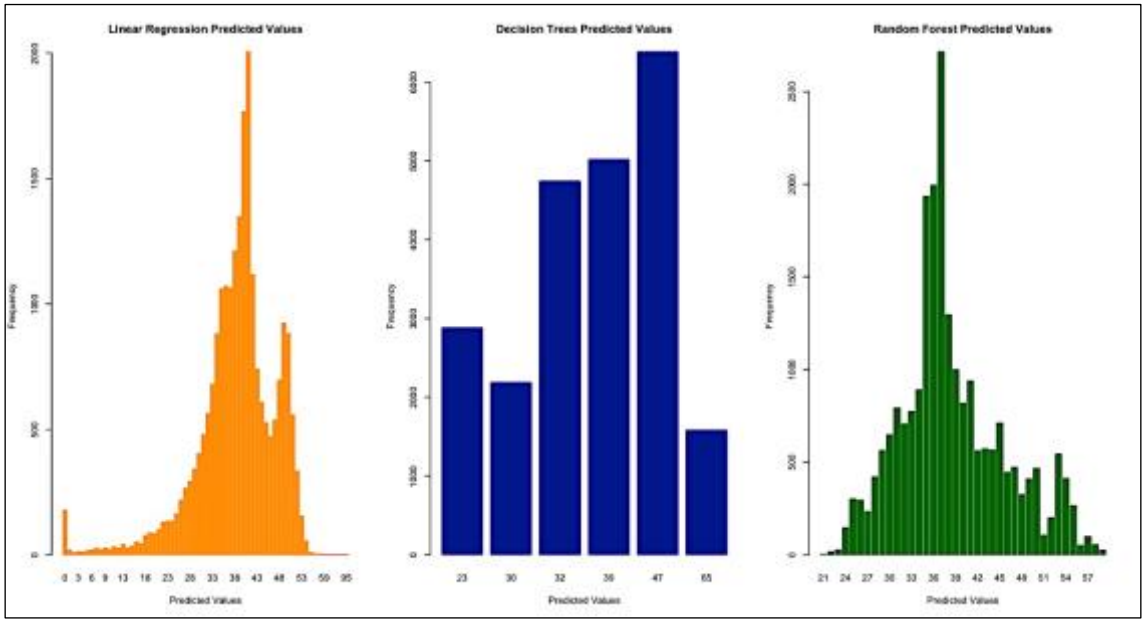

Figura 39: Distribuição dos valores de inventário previstos pelos modelos

Pela figura 39 é possível concluir que para o modelo de árvores de decisão os possíveis valores de inventário são demasiado limitados. Em relação aos modelos de regressão linear e random forests, o modelo de regressão linear é o que apresenta a maior quantidade de inventário de baixo valor e tendo em conta que os erros entre estes dois modelos estão muito próximos, então, é o modelo de regressão linear que deve ser aplicado.

6.3.4 Modelo questão de análise 5

Por forma a relembrar o que se pretende obter com a implementação deste modelo, a questão de análise é apresentada abaixo.

"Qual o valor monetário que é esperado obter na venda de produtos."

Pretende-se com esta questão de análise prever o valor a receber com a venda de produtos, quer sejam os que se encontram em inventário quer seja novos produtos.

Foi então definido que o valor monetário esperado obter é dado por:

```
Valor monetário = Preço * Quantidade Vendida
```

Devem ser criados dois modelos de previsão, um para o preço de vendas e outro para a quantidade a ser vendida. O modelo para esta ultima já foi efetuado na questão de análise 3, logo, resta efetuar o modelo de previsão de preço.

Foi necessário começar por decidir quais as variáveis a utilizar:

- *PriceReg* (variável a ser prevista)
- *MarketingType*
- *NewReleaseFlag*
- *StrengthFactor*
- *ReleaseYear*
- *ItemCount*

Para dar resposta à questão de análise foi, tal como nas questões anteriores, utilizado vários tipos de modelos de regressão (apresentados na tabela 12) por forma a obter o melhor.

6.3.4.1 Modelo de Regressão Linear

O modelo de regressão linear foi desenvolvido recorrendo à função abaixo, que utiliza o conjunto de variáveis indicadas.

```
linear_r_model_price <- lm(PriceReg ~ MarketingType + NewReleaseFlag + StrengthFactor + ReleaseYear + ItemCount, data = train_sales)
```

Tal como nos restantes modelos de regressão, foi aplicado este modelo sobre o conjunto de dados de teste e depois foi otimizado várias vezes pela eliminação de valores extremos recorrendo à distância de Cook.

A tabela 30 apresenta a comparação do valor dos erros e das métricas utilizadas para avaliar o modelo original e otimizado.

	MAE	RMSE	CORRELATION	R2
Original	53.28835512	81.00942579	0.17258600	0.02978593
Otimizado	53.12354768	81.01943055	0.17203071	0.02959456

Tabela 30: Erros MAE e RMSE e métricas Corr. e r-squared do modelo original e final

Como aconteceu com outros modelos, apesar de se ter desenvolvido um modelo otimizado, o mesmo é praticamente idêntico ao modelo original, muito provavelmente devido à presença de *outliers*.

6.3.4.2 Modelo de Árvores de Decisão

A construção deste modelo foi efetuada recorrendo à função abaixo que utiliza um método apropriado para problemas de regressão.

```
tree_r_model_price <- rpart(PriceReg ~ MarketingType + NewReleaseFlag + StrengthFactor + ReleaseYear + ItemCount, data = train_sales, method = "anova")
```

A figura 40 mostra a árvore gerada.

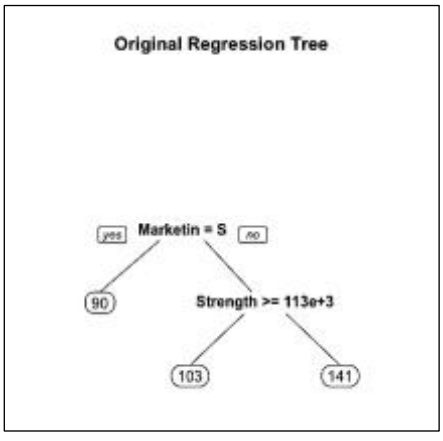

Figura 40: Árvore de decisão gerada

Observando a árvore acima conclui-se que não possui uma variação aceitável para possíveis preços de venda dos produtos. Foi então otimizado o modelo de forma a gerar mais possíveis valores para os preços de venda. A figura 41 apresenta a árvore final otimizada.

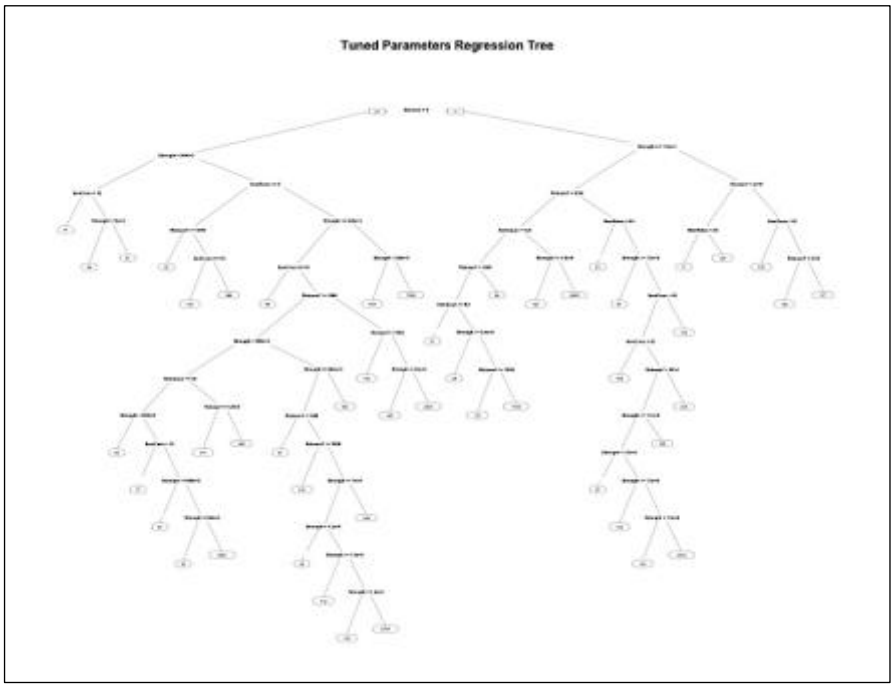

Figura 41: Árvore de decisão final (otimizada)

A árvore final possui um maior conjunto de possibilidade para os valores de venda dos produtos. Na tabela abaixo encontra-se os valores dos erros e das métricas utilizadas para comparar os dois modelos.

	MAE	RMSE	CORRELATION	R2
Original	53.38264760	81.27348225	0.15256124	0.02327493
Otimizado	52.21743793	83.37812021	0.14581073	0.02126077

Tabela 31: Erros MAE e RMSE e métricas Corr. e r-squared do modelo original e final

6.3.4.3 Modelo de Random Forests

O ultimo modelo desenvolvido é baseado em random forests. Foi construído pela função abaixo, com geração de 500 árvores.

```
randomF_r_model_price <- randomForest(PriceReg ~ MarketingType + NewReleaseFlag + StrengthFactor +
ReleaseYear + ItemCount, data = train_sales)
```

No anexo 5, figuras 52 e 53 pode ser consultada a representação visual dos erros do modelo bem como a importância das variáveis. A tabela 32 apresenta o valor dos erros e das métricas relevantes para avaliação do modelo.

MAE	RMSE	CORRELATION	R2
52.42570576	80.12034656	0.24618910	0.06060908

Tabela 32: Erros MAE e RMSE e métricas Corr. e r-squared do modelo

6.3.4.4 Escolha do Melhor Modelo

A tabela a baixo mostra a comparação dos valores de erros e métricas dos vários modelos desenvolvidos.

	MAE	RMSE	CORRELATION	R2
Regressão Linear (otimizado)	53.12354768	81.01943055	0.17203071	0.02959456
Árvores de Decisão (otimizado)	52.21743793	83.37812021	0.14581073	0.02126077
Random Forests	52.42570576	80.12034656	0.24618910	0.06060908

Tabela 33: Erros MAE e RMSE e métricas Corr. e r-squared dos vários modelos desenvolvidos

Com base nos valores da tabela o modelo de random forests é ligeiramente o melhor modelo dos três.

Uma vez desenvolvido o modelo para previsão do preço de venda e utilizando o modelo já desenvolvido anteriormente de previsão da quantidade de venda dos produtos, foi facilmente desenvolvida uma função que recebe os valores esperados de quantidade vendida de uma lista de produtos (*data frame*) e ainda os valores previstos de preço de venda e devolve o valor monetário esperado obter com cada produto.

7. Implementação

Uma vez que o desenvolvimento do presente trabalho assenta no âmbito de uma disciplina académica, não houve muita preocupação no desenvolvimento do trabalho para o desenvolver com o intuito de fazer parte de um programa de utilização por parte de uma entidade.

Contudo, no caso de isso acontecer, todo o código desenvolvido dos modelos seria facilmente integrado em funções que seriam executadas de forma oportuna. Para além do desenvolvimento dessas funções devia ser elaborado um mecanismo que consoante o volume de dados transacionados pela entidade de utilização do sistema, fosse efetuando a reamostragem dos dados com os dados novos e adaptando os modelos para que de forma dinâmica utilizassem os novos dados de forma apropriada para treino e teste, mantendo-se assim os modelos fiáveis ao longo do tempo.

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 42: 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 34: Output da função "*summary()*" sobre o *dataset*

Anexo 2 – Modelo Questão de Análise 2

- Matrizes de confusão e métricas obtidas para cada modelo de reamostragem do *dataset* para o modelo de regressão logística.

Método de reamostragem	Matriz de confusão	Precisão	Sensibilidade	Especificidade
Oversampling	Reference Prediction 0 1 0 13374 1323 1 5520 2581	0.6998	0.7078	0.6611
Undersampling	Reference Prediction 0 1 0 13389 1318 1 5505 2586	0.7007	0.7086	0.6624
Both Over. & Under.	Reference Prediction 0 1 0 13361 1318 1 5533 2586	0.6995	0.7072	0.6624
SMOTE	Reference Prediction 0 1 0 14750 1674 1 4144 2230	0.7448	0.7807	0.5712

Tabela 35: Matriz de confusão e métricas obtidas para cada modelo de reamostragem do *dataset*

- Representação gráfica dos erros do modelo original de random forests.

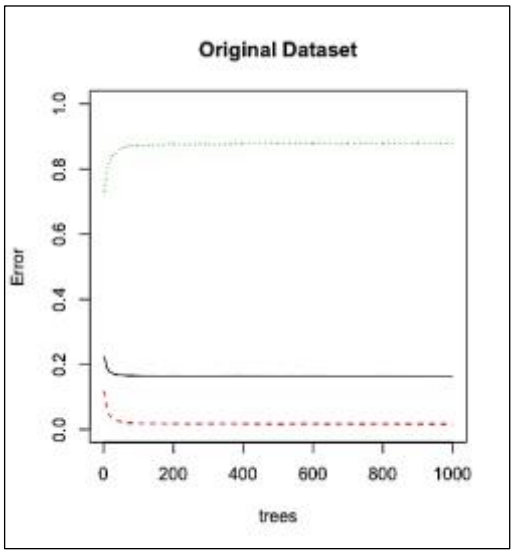

Figura 43: Representação dos erros do modelo original de random forests

- Representação gráfica dos erros dos modelos com dados reamostrados do modelo de random forests.

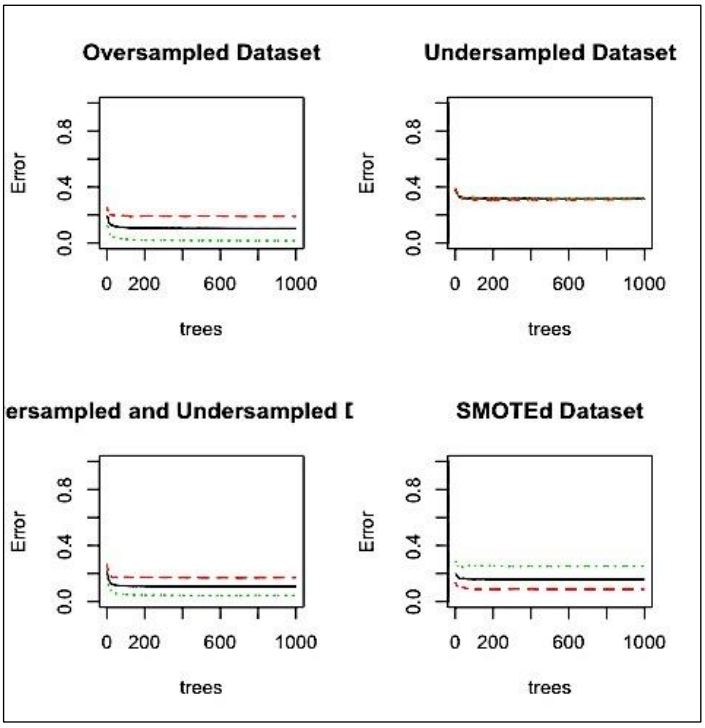

Figura 44: Representação dos erros de cada modelo de random forests com dados reamostrados

Anexo 3 – Modelo Questão de Análise 3

- Actual vs Predicted sobre os dados de treino e de teste do modelo de regressão linear.

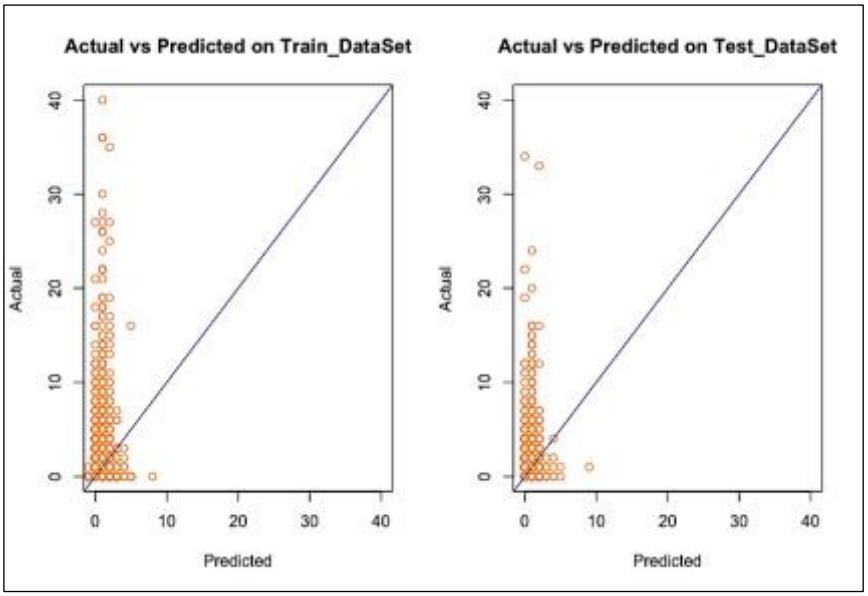

Figura 45: Actual vs Predicted sobre os dados de treino e de teste

- Árvore de decisão otimizada (intermédia).

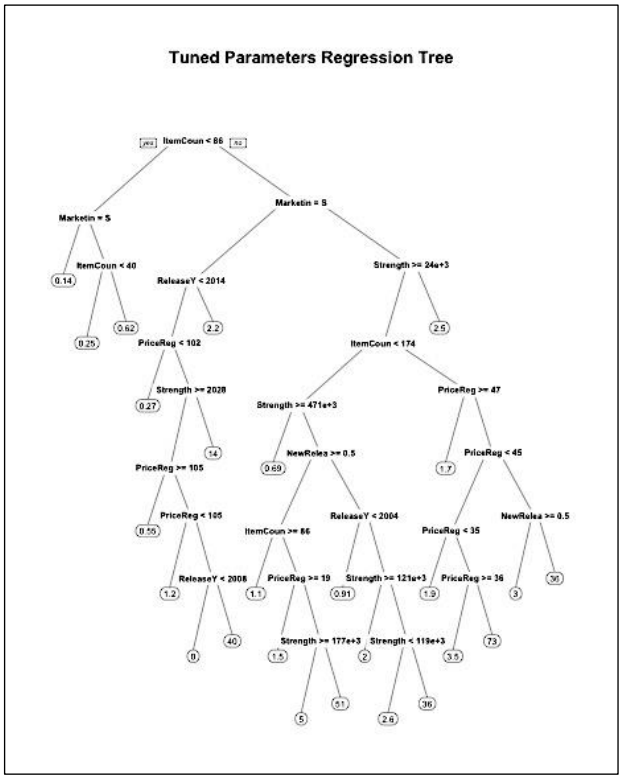

Figura 46: Árvore de decisão obtida pela otimização da árvore original

- Representação gráfica dos erros do modelo com random forests e importância das variáveis.

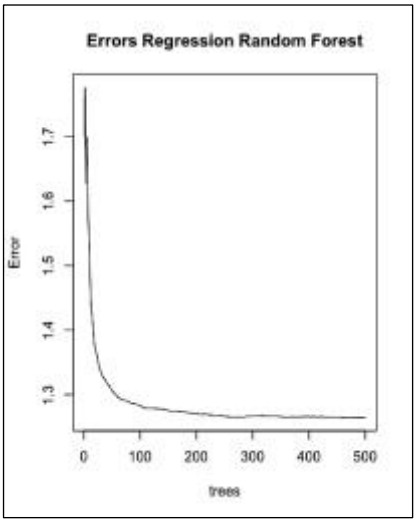

Figura 48: Representação dos erros do modelo de random forests

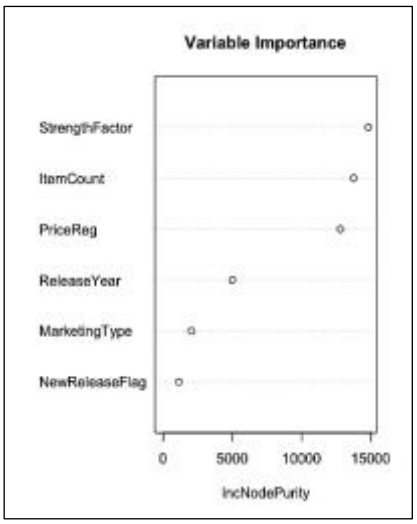

Figura 48: Representação da importância que cada variável tem sobre o modelo

Anexo 4 – Modelo Questão de Análise 4

- Representação gráfica dos erros do modelo original e otimizado de random forests e ainda representação da importância das variáveis para o modelo otimizado.

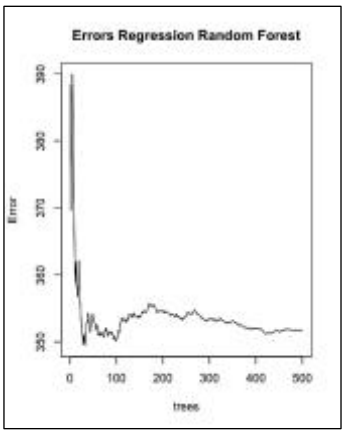

Figura 51: Erros modelo original

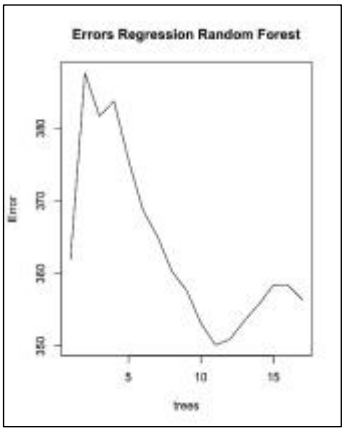

Figura 51: Erros modelo otimizado

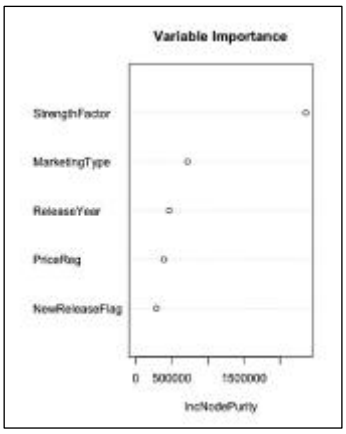

Figura 51: Importância das variáveis

Anexo 5 – Modelo Questão de Análise 5

- Representação gráfica dos erros do modelo de random forests e ainda representação da importância das variáveis.

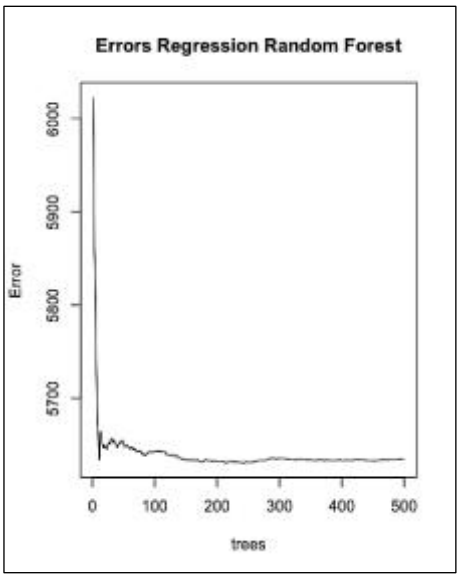

Figura 53: Erros do modelo

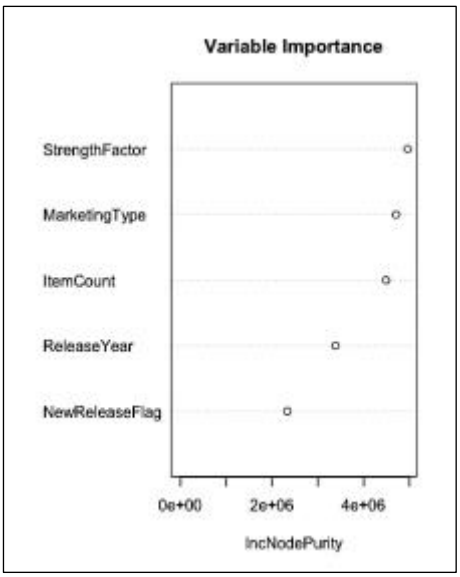

Figura 53: Importância das variáveis