Trabalho Final Grupo B2

Fonseca, A., Duarte, G. & Margarido, R.
2013170494, 2013155376, 2013145676

Resumo
Tendo em conta uma série de parâmetros considerados pertinentes, procurou-se implementar um método de previsão da compra ou venda de ações. Usaram-se os métodos de classificação aprendidos nas aulas durante o semestre para descrever as 'features' e identificar aquelas que melhor se relacionam com o output final. Criaram-se, então, dois classificadores distintos que foram, posteriormente, avaliados.

Conteúdo

Introdução

Um investidor da bolsa esporádico reuniu uma série de parâmetros que achava pertinentes e que geralmente usava para prever pontos de compra e venda de ações. Estes dados foram organizados numa base de dados, à qual se adicionou uma variável dicotómica de compra (0) e venda (1), com o intuito de se implementar um método de previsão da compra e venda de ações.

Quanto às variáveis, apenas é conhecido o seu nível de mensuração (V1 e V2 – nominal; V3 e V4 – ordinal; V5 e V6 – quantitativa). Existe ainda uma variável (R) na base de dados que corresponde aos rótulos para os quais se pretende implementar o modelo estatístico de classificação.

1. Métodos

1.1 Estudo das Variáveis

Para que as variáveis (V1, V2, V3, V4, V5 e V6) influenciem na decisão de compra/venda de ações elas têm de ser dependentes, ou seja, a mesma variável tem de ter um comportamento diferente na compra e venda de ações.

Começamos por obter as frequências relativas das diferentes variáveis para os valores de R (compra (0) e venda (1)): gráficos de barras para o caso das variáveis nominais e ordinais; gráficos de extremos e quartis (boxplot) para as variáveis quantitativas. Obtivemos também tabelas de dupla-entrada.

Efetuamos ainda vários testes para reforçar a análise dos gráficos e avaliar a correlação das diferentes variáveis com R. Para as variáveis nominais, V1 e V2, recorreu-se ao teste do qui-quadrado, para as variáveis ordinais, V3 e V4, usou-se o teste de Mann-Whitney. Para as variáveis quantitativas aplicou-se o teste de Shapiro-Wilk, seguido do teste de Mann-Whitney se a hipótese nula for rejeitada, caso contrário (hipótese nula não rejeitada) o teste de Shapiro-Wilk é seguido do teste de Levene e do teste t-student.

1.2 Classificadores

Procedeu-se à divisão das variáveis em 2 grupos: de treino e de teste. Utilizamos primeiro o grupo de treino para realizar o modelo

Através dos testes de independência realizados e considerando apenas as variáveis que apresentavam uma correlação com a variável dependente R, criamos um modelo de regressão logística usando a função glm. De seguida, recorremos ainda ao teste de Wald de modo a concluir acerca da capacidade descriminativa das variáveis, e ao teste de Hosmer-Lemeshow para avaliar o ajuste obtido. Fizemos também um outro modelo usando a função svm.

Por último, aplicamos os modelos ao grupo de teste, os quais avaliamos através da matriz confusão onde calculamos a exatidão, sensibilidade e especificidade.

2. Resultados

2.1 Estudo das Variáveis

2.1.1 Variáveis Nominais

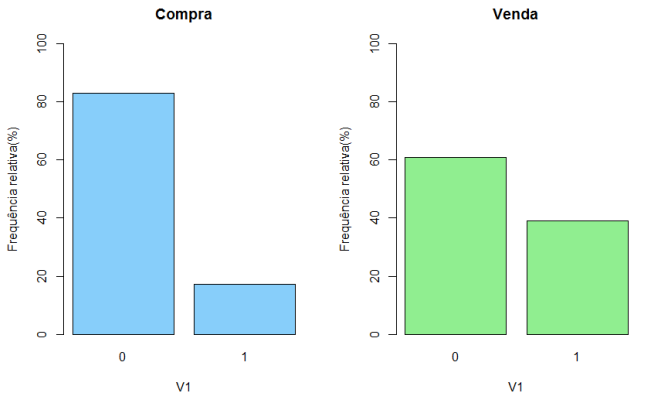

Figure 1. Representação gráfica das frequências relativas (%) de V1 em função dos rótulos "compra" e "venda" respetivamente.

A tabela referente à distribuição da variável é a seguinte:

	Compra	Venda
0	34	7
1	36	23

O valor de p associado ao teste deχ^2 para a variável V1 é 0.0332.

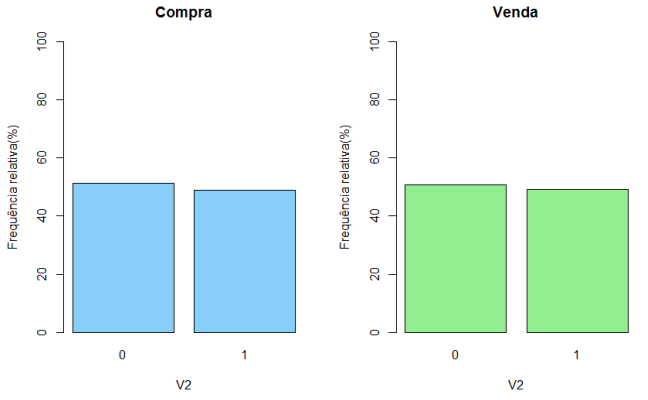

Figure 2. Representação gráfica das frequências relativas (%) de V2 em função dos rótulos "compra" e "venda" respetivamente.

A tabela referente à distribuição da variável é a seguinte:

	Compra	Venda
0	21	30
1	20	29

O valor de p associado ao teste de χ^2 para a variável V2 é 1.

2.1.2 Variáveis Ordinais

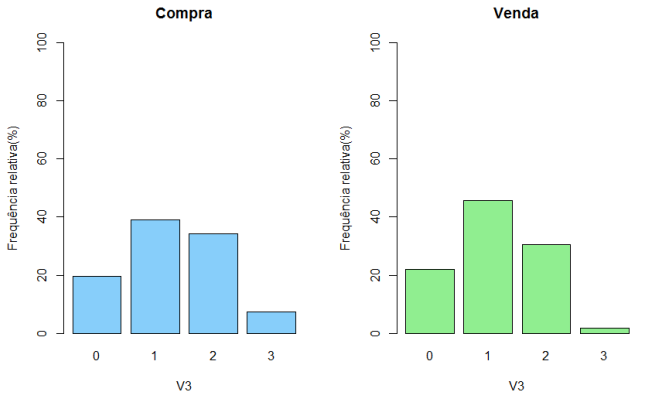

Figure 3. Representação gráfica das frequências relativas (%) de V3 em função dos rótulos "compra" e "venda" respetivamente.

A tabela referente à distribuição da variável é a seguinte:

	Compra	Venda
0	8	13
1	16	27
2	14	18
3	3	1

O valor de p associado ao teste de Wilcoxon para a variável V3 é 0.3404.

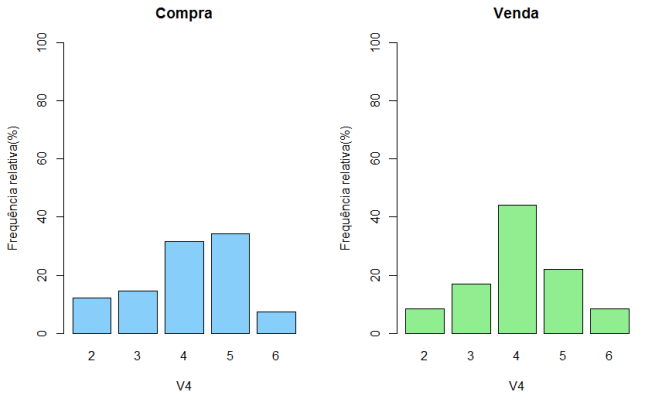

Figure 4. Representação gráfica das frequências relativas (%) de V4 em função dos rótulos "compra" e "venda" respetivamente.

A tabela referente à distribuição da variável é a seguinte:

	Compra	Venda
2	5	5
3	6	10
4	13	26
5	14	13
6	3	5

O valor de p associado ao teste de Wilcoxon para a variável V4 é 0.655.

2.1.3 Variáveis Quantitativas

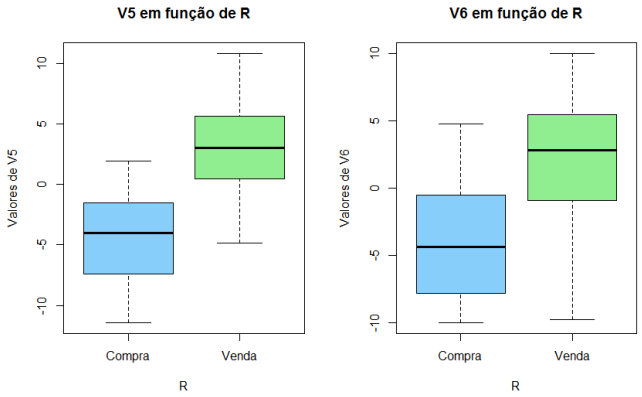

Figure 5. Boxplot dos valores de V5 e V6 em função dos rótulos "compra" e "venda" respetivamente.

Aplicando o teste de normalidade de Shapiro-Wilk à variável V5 em função do rótulo "compra" foram obtidos os valores W=0.96982 e p=0.3401. Respetivamente à mesma variável mas agora aplicada ao rótulo "venda" foram obtidos os valores W=0.98285 e p=0.5715. Recorrendo ao teste de Levene foram obtidos os seguintes valores F=0.0662 e p=0.7975. Por último, para o teste-T foram obtidos os valores t=-9.2698 e p=4.655E-15.

Aplicando o teste de normalidade de Shapiro-Wilk à variável V6 em função do rótulo "compra" foram obtidos os valores W=0.95743 e p=0.1278. Na mesma variável mas agora aplicada ao rótulo "venda" foram obtidos os valores W=0.95317 e p=0.02367. Recorrendo ao teste Wilcoxon rank sum foram obtidos os seguintes valores W=467 e p=1.991E-7.

2.2 Modelo Logístico

Aplicando o modelo glm às variáveis dependentes com recurso ao método de Wald obtemos:

	2.5%	97.5%
Intercept	1.1278251	10.242434
V1	-1.5072959	21.899889
V5	1.0997857	7.223672
V6	0.822784	3.697061

Considerando agora apenas V5 e V6 como variáveis discriminativas:

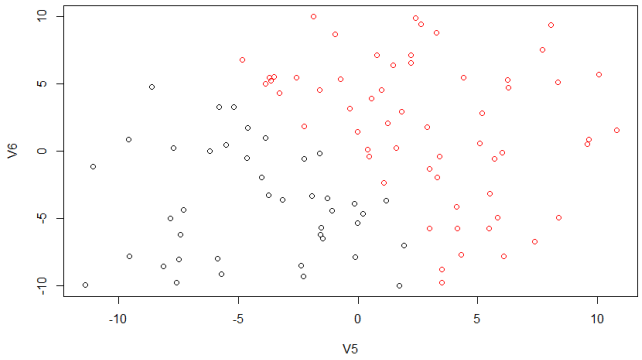

Figure 6. V6 em função de V5 em que o rótulo "compra" está a preto e o rótulo "venda" a vermelho (que aparenta formar dois grupos distinguíveis).

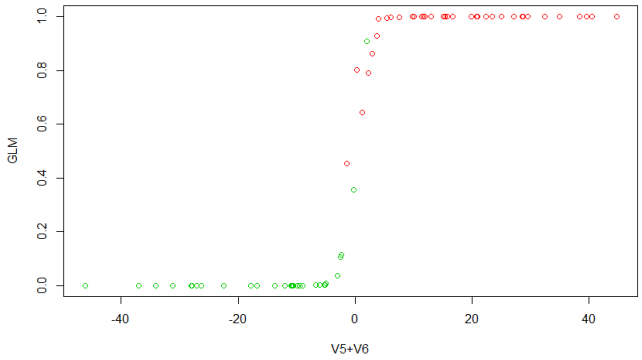

Figure 7. Gráfico obtido pela equação do modelo GLM

Um exemplo da tabela relativa ao intervalo de confiança para o modelo GLM é o seguinte:

Excluindo V1:

	2.5%	97.5%
Intercept	1.762261	13.69545
V5	1.304709	10.08727
V6	0.822784	5.60373

O modelo obtido através da regressão linear obedece à seguinte equação:

$$P = \frac{1}{1 + e^{4.769 + 2.985 * V5 + 1.697 * V6}}$$

Um exemplo de matriz de confusão associado a este modelo é:

	Compra	Venda
Compra	13	1
Venda	0	16

Após várias runs do programa foi calculado o valor médio da exatidão, sensibilidade e especificidade para o modelo GLM:

	Média%
Exatidão	93,6
Sensibilidade	91,7
Especificidade	96,3

Foi ainda calculado o valor médio de p no teste de Hosmer & Lemeshow: 0.9699

2.3 Modelo SVM

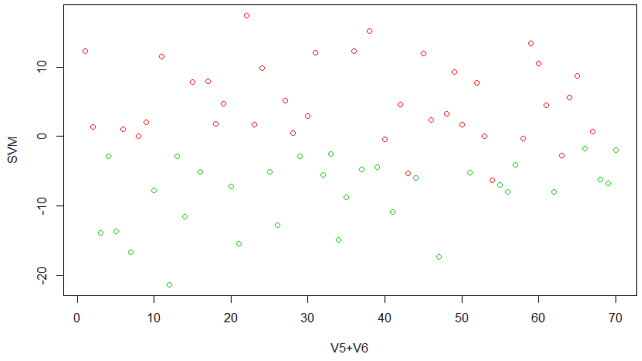

Figure 8. Distribuição de V5+V6 em função do modelo SVM

Um exemplo de matriz de confusão associado a este modelo é:

	Compra	Venda
Compra	10	0
Venda	3	17

Após várias runs do programa foi calculado o valor médio da exatidão, sensibilidade e especificidade para o modelo SVM:

	Média%
Exatidão	83,6
Sensibilidade	75,8
Especificidade	91,4

3. Discussão

3.1 Depêndencia das Variáveis em função de R

Para as variáveis nominais, a partir da observação da figura 1 e tabela correspondente é evidente uma diferença substancial entre as frequências relativas e os valores absolutos em função dos dois rótulos. Recorrendo ao teste qui quadrado foi possível verificar que, como p menor que α , pode-se rejeitar a hipótese nula (variáveis ser independentes) pelo que V1 é uma variável dependente. Estas diferenças não são tão evidentes na figura 2 e tabela correspondente sendo o p obtido para o teste qui quadrado superior a alfa pelo que V2 é uma variável independente.

Respetivamente às variáveis ordinais, embora as mesmas aparentam ter algumas diferenças na figura 3 e 4 que possam ser indicadores da dependência das variáveis, recorrendo ao teste Wilcoxon rank sum, para as variáveis V3 e V4 foram obtidos valores de p maiores que α pelo que, não se podendo rejeitar a hipótese nula que afirma a igualdade das medianas, ambas as variáveis são independentes.

Por último, para as variáveis quantitativas V5 e V6, é possível verificar na figura 5 que o 3^o quartil do rótulo "compra" corresponde a um valor menor que o 1^o quartil do rótulo "venda" e que na figura 6 estes valores são sensivelmente iguais. De forma a avaliar a potencial dependência das variáveis, recorrendo a um teste de normalidade de Shapiro-Wilk para V5, foram obtidos valores de p maior que α para ambos os rótulos pelo que ambas seguem uma distribuição normal. Pelo teste de Levene foi obtido um valor de p maior que α pelo que se verifica a homocedasticidade da variável. Com o teste-T o valor de p obtido foi inferior a α pelo que se rejeita a hipótese nula que afirma a igualdade das médias pelo que V5 é uma variável dependente. Por oposição, a variável V6 em função do rótulo "venda" apresentou um valor de pmenor que α pelo que não se verifica uma distribuição normal sendo portanto necessário recorrer ao teste wilcoxon rank sum para o qual se obteve p menor α, rejeitando-se a hipótese nula sendo portanto as medianas diferentes. Deste modo V6 é dependente.

3.2 GLM & SVM

Considerando as variáveis dependentes como variáveis discriminativas, aplicando o modelo glm ao conjunto V1+V5+V6, pelo método de Wald foi possível verificar a presença do 0 no intervalo de confiança da variável V1 pelo que a mesma não é relevante para a discriminação do modelo.

Pela observação da figura 8 de V6 em função de V5 é possível observar dois grupos distinguíveis pelo que se aplicará novamente o modelo glm e posteriormente o svm mas agora ao conjunto de variáveis V5+V6.

Recorrendo novamente ao método de Wald é possível agora verficiar a inexistência do 0 nos intervalos de confiança de V5 e V6.

Para a criação de ambos os modelos (GLM e SVM) recorreu-se à criação de dois grupos: um de treino e outro de teste. Como estes grupos são criados aleatoriamente fazendo um sample dos dados é importante a repetição do programa para que se possa registar um número satisfatório de dados e fazer a sua média para eliminar esta aleatoriedade.

Uma vez que no caso em questão estamos a tratar da compra e venda de ações em bolsa não podemos apenas olhar

para um destes parâmetros mas sim para os 3 uma vez que um erro se pode revelar não só numa perda avultada de capital mas também numa perda de potencial lucro. É importante que o modelo não indique a venda ou compra de ações em situações que não sejam as mais vantajosas para o investidor.

Assim sendo e olhando para as tabelas apresentadas é evidente que o modelo GLM apresenta melhores resultados que o SVM (apesar de este ser ajustável nunca foi capaz de obter valores melhores que o GLM), sem que contudo apresente um overfitting aos dados apresentados (o que resultaria na perda de capacidade preditiva por parte do classificador).

Ao ter em conta o valor elevado do teste de Homer-Lemeshow também conferimos o bom ajuste do modelo à situação apresentada.

4. Conclusão

O trabalho acima apresentado tem como objetivo a aplicação de conhecimentos e técnicas ensinadas ao longo do semestre para elaboração de um classificador eficiente para a situação apresentada. Para o fazer recorremos a dois modelos distintos GLM e SVM e concluimos que o GLM apresenta melhores resultados para o caso em estudo.

Na situação económica que vivemos é importante que o classificador avalia o mercado de forma meticulosa dada as flutuações esporádicas do mesmo. Sendo assim e após a averiguação de quais as melhores variáveis para a elaboração do dito classificador apresentamos um modelo de regressão logística com valores de exatidão, especificidade e sensibilidade bastante promissores.

O modelo é portanto extremamente capaz de informar o investidor acerca das decisões que o mesmo tem de tomar em relação à compra e venda de ações. De referir que devido à aplicabilidade do mesmo este servirá para analisar grandes volumes de informação num curto de espaço de tempo devolvendo ao utilizador a informação relevante no final.