Trabalho final de disciplina Estatística – Uni7
Curso Pós Graduação em Ciência de Dados – Turma 06
Equipe:
Francisco Flávio Cardoso Gomes
Jean Carlos Maia e Silva
Israel Portela

QUESTÃO 1 - (3 pontos):

Uma empresa seguradora deseja prever com maior segurança as possibilidades de indenização das apólices de seguro de automóveis quando assume um novo contrato. Assim, através do levantamento de registros anteriores, obteve os seguintes dados:

Esses casos foram divididos em três grupos: (1) Baixo Risco, (2) Médio Risco e (3) Alto Risco. As variáveis consideradas representativas foram "tempo de habilitação", "número de multas" sofridas desde que obteve sua habilitação e estado civil/existência ou não de filhos, assumindo os valores de 1 para solteiro, 2 para casado e 3 com filhos.

Observação	Grupo	Tempo de habilitação	Solteiro (1) Casado (2) Filhos? (3)	Número de multas
1	1	20	3	1
2	1	21	3	0
3	1	25	3	2
4	1	25	2	3
5	1	18	2	2
6	1	23	1	2
7	2	9	3	6
8	2	12	2	4
9	2	15	1	3
10	2	14	2	2
11	2	15	1	5
12	2	10	3	5
13	2	8	2	4
14	3	7	2	13
15	3	11	1	15
16	3	10	2	9
17	3	7	2	6
18	3	9	1	10
19	3	1	3	8
20	3	3	1	5

Pede-se:
1) Usando o método da distância de Mahalanobis, estabeleça o grau de risco de cada novo potencial cliente cujos dados são:

2) Indique a probabilidade em que cada um dos novos clientes pode pertencer ao Grupo Correspondente e cria uma Regra de Negócio diante dessa problemática. Aponte ainda qual dessas três variáveis é a que mais discrimina a classificação dos Grupos.

Utilizamos o Software SPSS da IBM para realizar os cálculos

Os Resultados são mostrados abaixo.

Grupo	TempoHab	Status	Multas	Aleat	Dis_1	Dis1_1	Dis2_1	Dis1_2	Dis2_2	Dis3_2
1	20	3	1	1,00	1	2,96778	-,34727	,98959	,01041	,00000
1	21	3	0	,00	1	3,59577	-,46452	,99869	,00131	,00000
1	25	3	2	,00	1	3,99486	,60660	,99986	,00014	,00000
1	25	2	3	,00	1	3,64272	,86329	,99961	,00039	,00000
1	18	2	2	,00	1	2,06395	-,36945	,80381	,19619	,00000
1	23	1	2	,00	1	3,44317	,32773	,99881	,00119	,00000
2	9	3	6	1,00	2	-1,82723	-,59762	,00000	,80838	,19162
2	12	2	4	1,00	2	-,29541	-,69269	,00090	,99861	,00049
2	15	1	3	,00	2	,88427	-,53107	,05738	,94262	,00000
2	14	2	2	1,00	2	,96057	-,92719	,05473	,94526	,00000
2	15	1	5	,00	2	,17998	-,01769	,00793	,99194	,00012
2	10	3	5	1,00	2	-1,19924	-,71487	,00004	,98267	,01729
2	8	2	4	,00	2	-1,39879	-1,25043	,00001	,97531	,02468
3	7	2	13	,00	3	-4,84392	,92033	,00000	,00001	,99999
3	11	1	15	,00	3	-4,44483	1,99145	,00000	,00002	,99998
3	10	2	9	1,00	3	-2,60782	,31188	,00000	,08238	,91762
3	7	2	6	,00	3	-2,37892	-,87649	,00000	,36907	,63093
3	9	1	10	1,00	3	-3,23580	,42913	,00000	,00662	,99338
3	1	3	8	,00	3	-4,73828	-1,19973	,00000	,00006	,99994
3	3	1	5	1,00	3	-3,13016	-1,69092	,00000	,05340	,94660
.	8	1	3	.	2	-1,04664	-1,50712	,00004	,99495	,00502
.	2	1	6	.	3	-3,75814	-1,57367	,00000	,00417	,99583
.	8	2	2	.	2	-,69450	-1,76381	,00010	,99890	,00100
.	20	2	4	.	1	1,91135	,42280	,81909	,18091	,00000
.	7	3	8	.	3	-3,08321	-,36312	,00000	,02272	,97728
.	15	3	1	.	2	1,58856	-1,04445	,31663	,68337	,00000
.	6	2	10	.	3	-4,06334	,01083	,00000	,00034	,99966
.	3	3	5	.	3	-3,13016	-1,69092	,00000	,05340	,94660

Resultados do teste

M de Box		8,354
Z	Aprox.	,889
	df1	6
	df2	578,405
	Sig.	,502

Testa hipótese nula de matrizes de covariâncias de população igual.

Resumo de processamento de caso de análise

Casos não ponderados	N	Porcentagem
Válido	12	60,0

Válido		12	60,0
Excluídos	Códigos de grupo omissos ou fora do intervalo	0	,0
	Pelo menos uma variável discriminante omisso	0	,0
	Códigos de grupo omissos ou fora do intervalo e pelo menos uma variável discriminadora omisso	0	,0
	Não selecionado	8	40,0
	Total	8	40,0
Total		20	100,0

Variáveis Inseridas/Removidas[a,b,c,d]

				Mín. Quadrado D			
					F exato		
Etapa	Inseridas	Estatística	Entre Grupos	Estatística	df1	df2	Sig.
1	TempoHab	2,879	2 e 3	4,936	1	9,000	,053
2	Multas	12,597	1 e 2	10,497	2	8,000	,006

Em cada passo, a variável que maximiza a distância de Mahalanobis entre os dois grupos mais próximos é inserida.

Resultados da classificação[a,b,d]

				Associação ao grupo prevista			
			Grupo	1	2	3	Total
Casos selecionados	Original	Contagem	1	5	0	0	5
			2	0	3	0	3
			3	0	0	4	4
		%	1	100,0	,0	,0	100,0
			2	,0	100,0	,0	100,0
			3	,0	,0	100,0	100,0
	Com validação cruzada[c]	Contagem	1	5	0	0	5
			2	0	3	0	3
			3	0	1	3	4
		%	1	100,0	,0	,0	100,0
			2	,0	100,0	,0	100,0
			3	,0	25,0	75,0	100,0
Casos não selecionados	Original	Contagem	1	1	0	0	1
			2	0	4	0	4
			3	0	0	3	3
		%	1	100,0	,0	,0	100,0
			2	,0	100,0	,0	100,0
			3	,0	,0	100,0	100,0

a. 100.0% de casos agrupados originais selecionados classificados corretamente.

b. 100.0% de casos agrupados originais não selecionados classificados corretamente.

Resposta à Questão 1/1

Observação	T. Hab	Status	Multas	Grupo	Risco	Probababilidades			P. de Corte	GRUPO ATRIBUIDO
1	8	1	3	2	Médio	0.000	0.995	0.005	0.547	2
2	2	1	6	3	Alto	0.000	0.004	0.996	0.548	3
3	8	2	2	2	Médio	0.000	0.999	0.001	0.549	2
4	20	2	4	1	Baixo	0.819	0.181	0.000	0.450	1
5	7	3	8	3	Alto	0.000	0.023	0.977	0.538	3
6	15	3	1	2	Médio	0.317	0.683	0.000	0.376	2
7	6	2	10	3	Alto	0.000	0.000	1.000	0.550	3
8	3	3	5	3	Alto	0.000	0.053	0.947	0.521	3

Resposta à Questão 1/2

O SPSS descartou o atributo Estado Civil, considerando apenas Tempo de Habilitação e Multas

As Probabilidades estão mostradas acima, juntamente com o Resultado da Regra de negócio

Para a obtenção da Regra de Negócio foram usados os cálculos de probabilidade para cada grupo para que seja extraído a maior probabilidade além de atribuir um peso de 55% sobre mesmos, dessa forma verifica-se se esse novo Ponto de Corte é menor que a maior probabilidade e maior que a menor probabilidade.

A variavel Multas possue menor discriminação por ter a uma significancia de 0,006.

QUESTÃO 2 - (3 Pontos)

O controlador da XPTO deseja determinar a influência das variáveis mão de obra (MO) e energia elétrica (EE) nos custos totais de fabricação (CTF) de seus produtos. Para isso, fez um levantamento dos valores destas variáveis nos últimos 24 meses e os resultados são mostrados na tabela, onde XXX representa os três últimos algarismos do seu número de matrícula na UNI7 (escolha a matrícula de um dos alunos da equipe, se for o caso).

Faça as análises de regressão simples e múltipla e determine o modelo de previsão para os CTF em função de MO e EE. Calcule os CTF para os seguintes valores:

Dados da questão

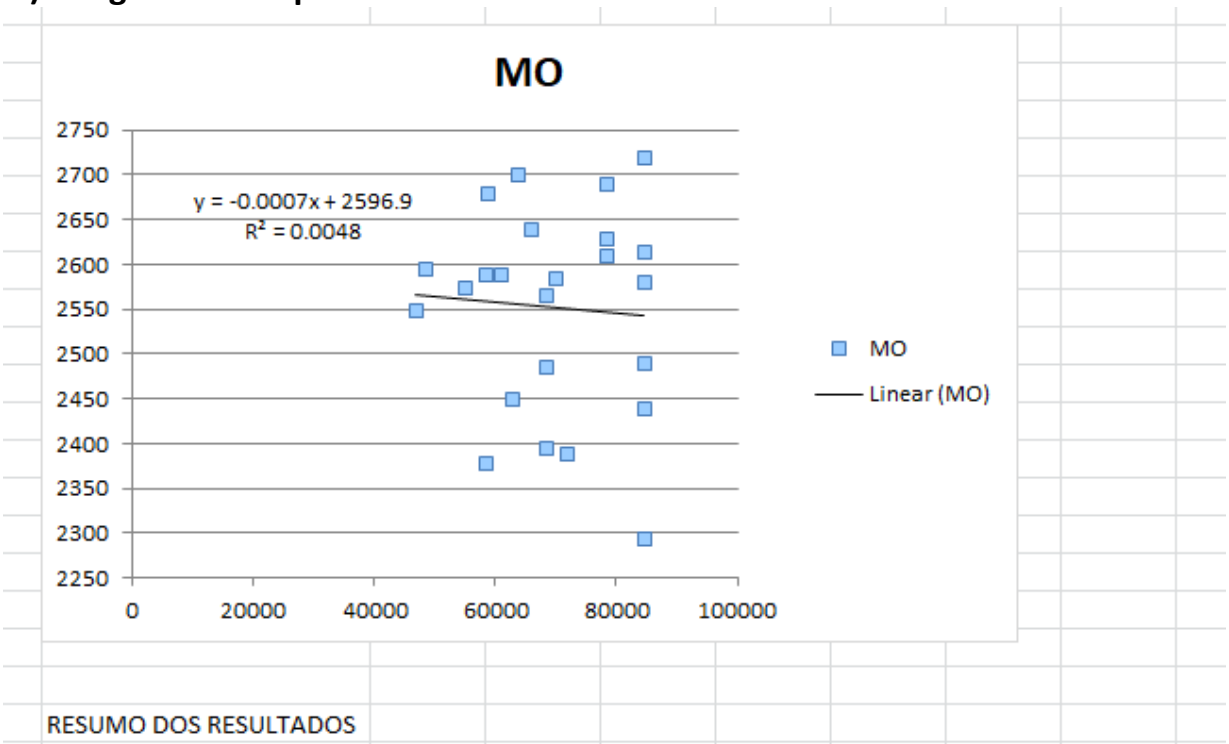
Meses	CTF	MO	EE
1	5XXX8	2378	980
2	XXX34	2295	945
3	62XXX	2450	930
4	6XXX7	2487	995
5	71XXX	2390	985
6	46XXX	2550	1010
7	XXX67	2440	998
8	5XXX9	2590	1025
9	7XXX3	2610	1100
10	54XXX	2575	1045
11	XXX98	2490	1038
12	XXX75	2580	1095
13	6XXX3	2395	1150
14	65XXX	2640	1030
15	4XXX9	2595	1085
16	XXX68	2720	1175
17	7XXX0	2690	1190
18	6XXX8	2565	1165
19	69XXX	2585	1200
20	XXX49	2615	1195
21	60XXX	2590	1210
22	7XXX1	2630	1189
23	58XXX	2680	1205
24	63XXX	2700	1200

Dados Utilizados (Matricula final 845)

	B	C	D
	CTF	MO	EE
	58458	2378	980
	84534	2295	945
	62845	2450	930
	68457	2487	995
	71845	2390	985
	46845	2550	1010
	84567	2440	998
	58459	2590	1025
	78453	2610	1100
	54845	2575	1045
	84598	2490	1038
	84575	2580	1095
	68453	2395	1150
	65845	2640	1030
	48459	2595	1085
	84568	2720	1175
	78450	2690	1190
	68458	2565	1165
	69845	2585	1200
	84549	2615	1195
	60845	2590	1210
	78451	2630	1189
	58845	2680	1205
	63845	2700	1200

Utilizamos o Excel para a Análise de Regresão

1) - Regressão Simples com Variável MO



RESUMO DOS RESULTADOS

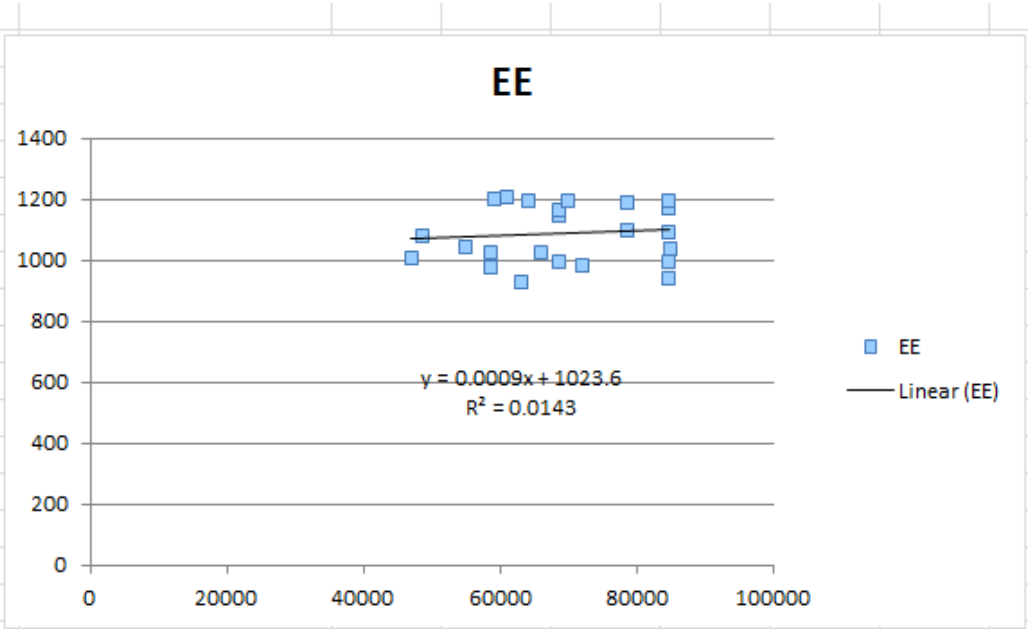
Estatística de regressão	
R múltiplo	0.06919
R-Quadrado	0.004787 (BAIXISSIMA CORRELAÇÃO)
R-quadrado ajustado	-0.04045
Erro padrão	12224.78
Observações	24

ANOVA

	gl	*MQ*	*F*	*e significação*	
Regressão	1	15815373	0.105827	0.748019	Significância > 0.05
Resíduo	22	1.49E+08			(SEM SIGNIFICÂNCIA)
Total	23				

	Coeficiente	*Stat t*	*valor-P*	*% inferior*	*% superior*	*iferior 95.0*	*perior 95.0%*
Interseção	88316.18	1.529168	0.140475	-31459.1	208091.5	-31459.1	208091.5
MO	-7.35621	-0.32531	0.748019	-54.2524	39.53998	-54.2524	39.53998

2) - Regressão Simples com Variável EE



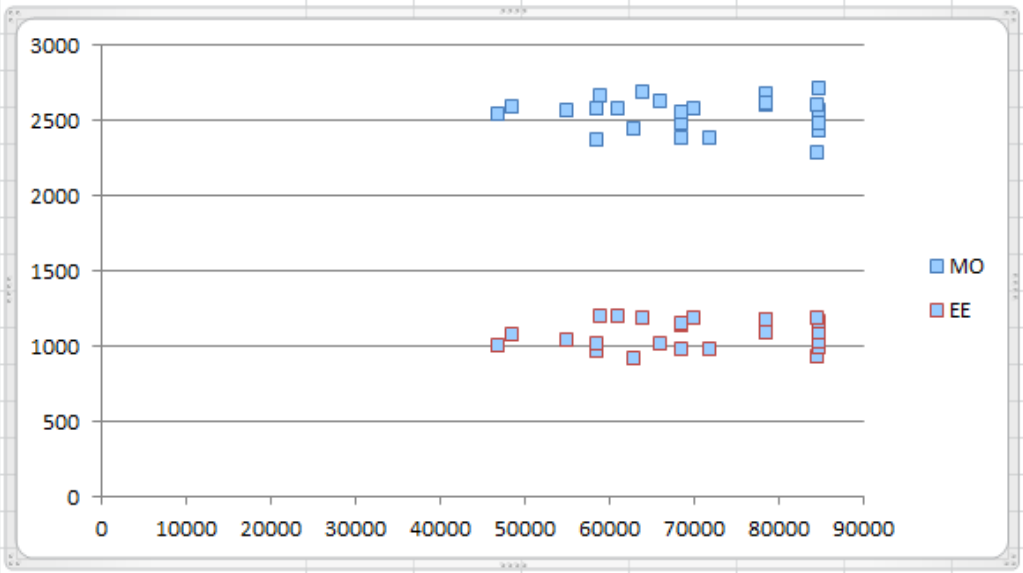
Estatística de regressão	
R múltiplo	0.940305
R-Quadrado	0.884173 (ALTA CORRELAÇÃO)
R-quadrado ajustado	0.878908
Erro padrão	4170.499
Observações	24

ANOVA

	gl	*SQ*	*MQ*	*F*	*e significação*	
Regressão	1	2.92E+09	2.92E+09	167.9384	8.96E-12	Significância < 0.05
Resíduo	22	3.83E+08	17393066			(BOA SIGNIFICÂNCIA)
Total	23	3.3E+09				

	Coeficiente	erro padrã	Stat t	valor-P	% inferior	% superio	ferior 95.0	or 95.0
Interseção	-60330.2	10058.07	-5.9982	4.89E-06	-81189.4	-39471.1	-81189.4	#
EE	119.2433	9.201507	12.95911	8.96E-12	100.1605	138.3261	100.1605	#

3) - Regressão Múltipla



RESUMO DOS RESULTADOS

Estatística de regressão	
R múltiplo	0.943606
R-Quadrado	0.890392 (ALTA CORRELAÇÃO)
R-quadrado ajustado	0.879953
Erro padrão	4152.471
Observações	24

ANOVA

	gl	SQ	MQ	F	e significação	
Regressão	2	2.94E+09	1.47E+09	85.29564	8.29E-11	Significância < 0.05
Resíduo	21	3.62E+08	17243014			(BOA SIGNIFICÂNCIA)
Total	23	3.3E+09				

	Coeficiente	erro padrã	Stat t	valor-P	% inferior	% superio	ferior 95.0	perior 95.0%
Intersecão	-39047.5	21919.46	-1.78141	0.089315	-84631.5	6536.466	-84631.5	6536.466

MO	-8.38458	7.681461	-1.09153	0.287403	-24.3591	7.589893	-24.3591	7.589893
EE	119.3461	9.162214	13.0259	1.58E-11	100.2922	138.4	100.2922	138.4

4) - Cálculo para novos valores - Utilizando as duas variáveis

Meses	CTF	MO	EE
25	72024	2695	1120
26	80712	2584	1185
27	78820	2710	1178
28	82084	2705	1205
29	80807	2715	1195

5) - Cálculo para novos valores - Utilizando apenas EE

Meses	CTF		EE
25	73222		1120
26	80973		1185
27	80138		1178
28	83358		1205
29	82165		1195

6) - Cálculo dos erros

Meses	CTF	MO	EE	Predito-2 Variáveis	ErroAbs	Predito - EE	ErroAbs
1	58458	2378	980	57973	485	56528	1930
2	84534	2295	945	347634	263100	52355	32179
3	62845	2450	930	71944	9099	50566	12279
4	68457	2487	995	79702	11245	58317	10140
5	71845	2390	985	78508	6663	57124	14721
6	46845	2550	1010	81492	34647	60105	13260
7	84567	2440	998	80060	4507	58675	25892
8	58459	2590	1025	83282	24823	61894	3435
9	78453	2610	1100	92233	13780	70837	7616
10	54845	2575	1045	85669	30824	64279	9434
11	84598	2490	1038	84834	236	63444	21154
12	84575	2580	1095	91636	7061	70241	14334
13	68453	2395	1150	98200	29747	76800	8347
14	65845	2640	1030	83879	18034	62490	3355
15	48459	2595	1085	90443	41984	69049	20590
16	84568	2720	1175	101184	16616	79781	4787
17	78450	2690	1190	102974	24524	81569	3119
18	68458	2565	1165	99991	31533	78588	10130
19	69845	2585	1200	104168	34323	82762	12917
20	84549	2615	1195	103571	19022	82165	2384
21	60845	2590	1210	105361	44516	83954	23109
22	78451	2630	1189	102855	24404	81450	2999

23	58845	2680	1205	104764	45919	83358	24513
24	63845	2700	1200	104168	40323	82762	18917
				Erro Absoluto =>	777417		301540
				Erro Padrão=>	10997		2300

Resposta

Analisando a regressão simples, verifica-se uma fraca correlação entre MO e CTF e uma forte correlação entre EE e CTF.

Na Regressão Múltipla, isso se confirma quando vemos a fraca significância da variável MO e uma forte significância da variável EE

O melhor ajuste se mostra na regressão simples com a variável EE, como mostra o cálculo do Erro Padrão, na tabela acima

O Resultado para a regressão simples (EE) e múltipla (MO e EE) são mostrados nas 2 tabelas anteriores

QUESTÃO 3 - (4 Pontos)

Desenvolva uma aplicação simulada em que a Regressão Logística seja a técnica de Análise Multivariada de Dados mais adequada para encontrar a Probabilidade do Evento [p(evento)] ocorrer. Indique a variável dependente e, pelo menos, três variáveis independentes estatisticamente significantes que possam apontar o(a) aumento/diminuição de chance de identificar a probabilidade do p(evento). Indique todos os testes adequados, apresente a Tabela de Classificação com VP, FP, FN e VN e as medidas de desempenho. Crie uma regra de classificação e apresente a solução da Regra de Negócio da simulação.

Conjunto de dados em anúncios de mídia social que descrevem se os usuários compraram um produto clicando nos anúncios exibidos a eles.

Dataset fonte: https://www.kaggle.com/akram24/social-network-ads

Variáveis independentes: Sexo, Idade e Salário
Purchased - 0 - Não comprou

Purchased - 1 - Comprou

User ID	Gender	Age	EstimatedSalary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0
15728773	Male	27	58000	0
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	0
15727311	Female	35	65000	0
15570769	Female	26	80000	0
15606274	Female	26	52000	0
15746139	Male	20	86000	0
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1
15617482	Male	45	26000	1
15704583	Male	46	28000	1
15621083	Female	48	29000	1
15649487	Male	45	22000	1
15736760	Female	47	49000	1
15714658	Male	48	41000	1
15599081	Female	45	22000	1
15705113	Male	46	23000	1
15631159	Male	47	20000	1
15792818	Male	49	28000	1
15633531	Female	47	30000	1
15744529	Male	29	43000	0
15669656	Male	31	18000	0
15581198	Male	31	74000	0
15729054	Female	27	137000	1
15573452	Female	21	16000	0
15776733	Female	28	44000	0
15724858	Male	27	90000	0

Utilizamos o SPSS para realizar a Regressão Logística,cujos resultados abaixo:

Variáveis na equação

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Etapa 1[a]	Gender(1)	-0.334	0.305	1.196	1	0.274	0.716	0.394	1.303
	Age	0.237	0.026	80.710	1	0.000	1.267	1.204	1.335
	EstimatedSalary	0.000	0.000	44.336	1	0.000	1.000	1.000	1.000
	Constante	-12.450	1.309	90.435	1	0.000	0.000		

a. Variável(is) inserida(s) no passo 1: Gender, Age, EstimatedSalary.

Análise do Modelo	
	%
Acurácia	85%
Precisão	92%
Recall	86%

Tabela de Classificação[a] - Confusão

			Previsto		
			Purchased		
	Observado		0	1	Porcentagem correta
Etapa 1	Purchased	0	237	20	92.2
		1	39	104	72.7
	Porcentagem global				85.3

a. O valor de recorte é ,500

P(Evento) = 1/1+2,7182^-(-12,45+(-0,334Gender)+(0,237Age)+(0,0Salary))

Calculadora

Regra de Negócio definida: Igual ou superior a 60%, ou seja, direcionar as campanhas para quem estiver dentro deste parâmetro

2.7182 e

	Sexo	Idade	Salário
Entre c/valor para simular	0	45	15000

Percentual
14%