BIX TECNOLOGIA

DESAFIO
CIENTISTA DE DADOS

RODRIGO VIEIRA

BIX TECNOLOGIA

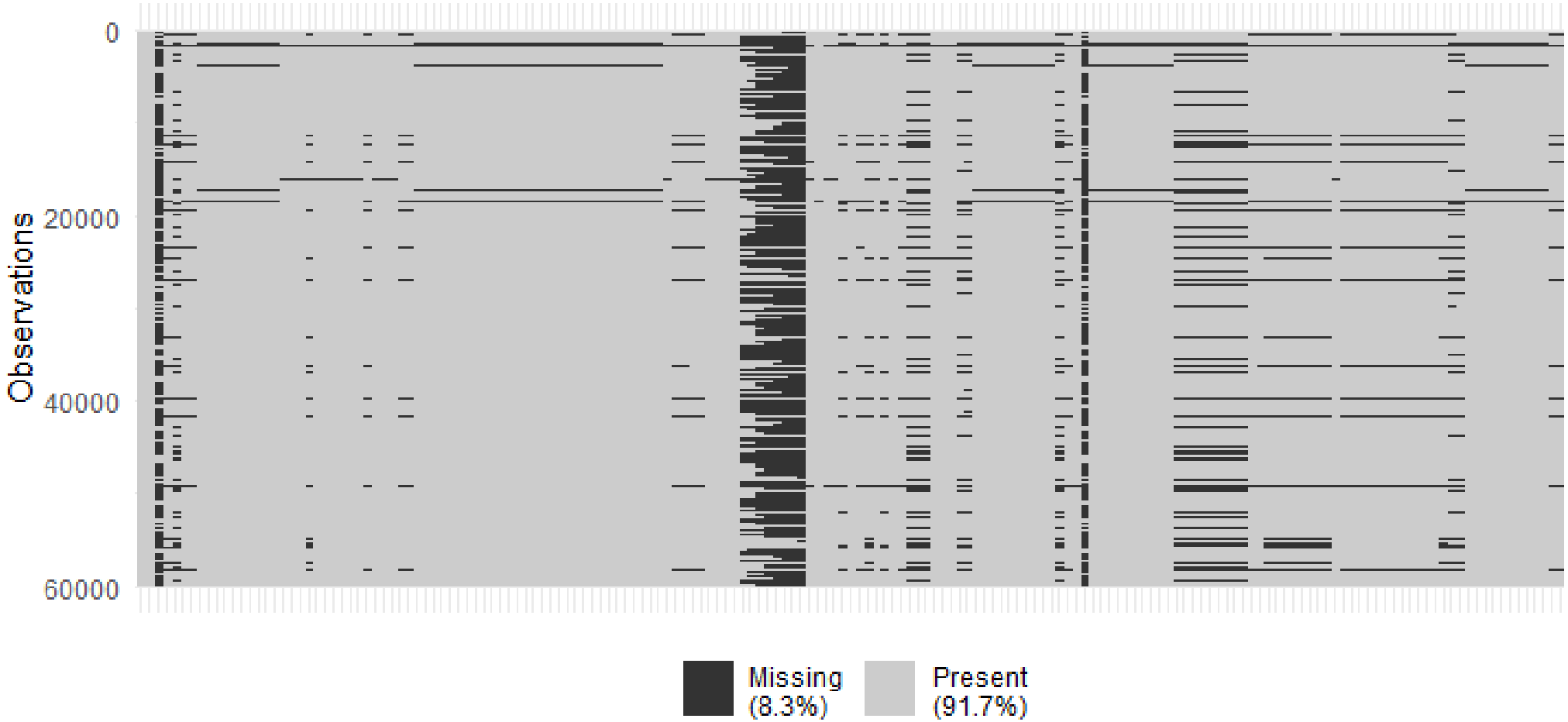
TOPICOS DA APRESENTAÇÃO

- PRINCIPAIS AÇÕES DE LIMPEZA E TRANSFORMAÇÃO
- ANÁLISE EXPLORATÓRIA
- ALGORITMOS APLICADOS
- RESULTADOS ALCANÇADOS
- SUGESTÕES E RECOMENDAÇÕES

BIX TECNOLOGIA

AÇÕES DE LIMPEZA E TRANSFORMAÇÃO

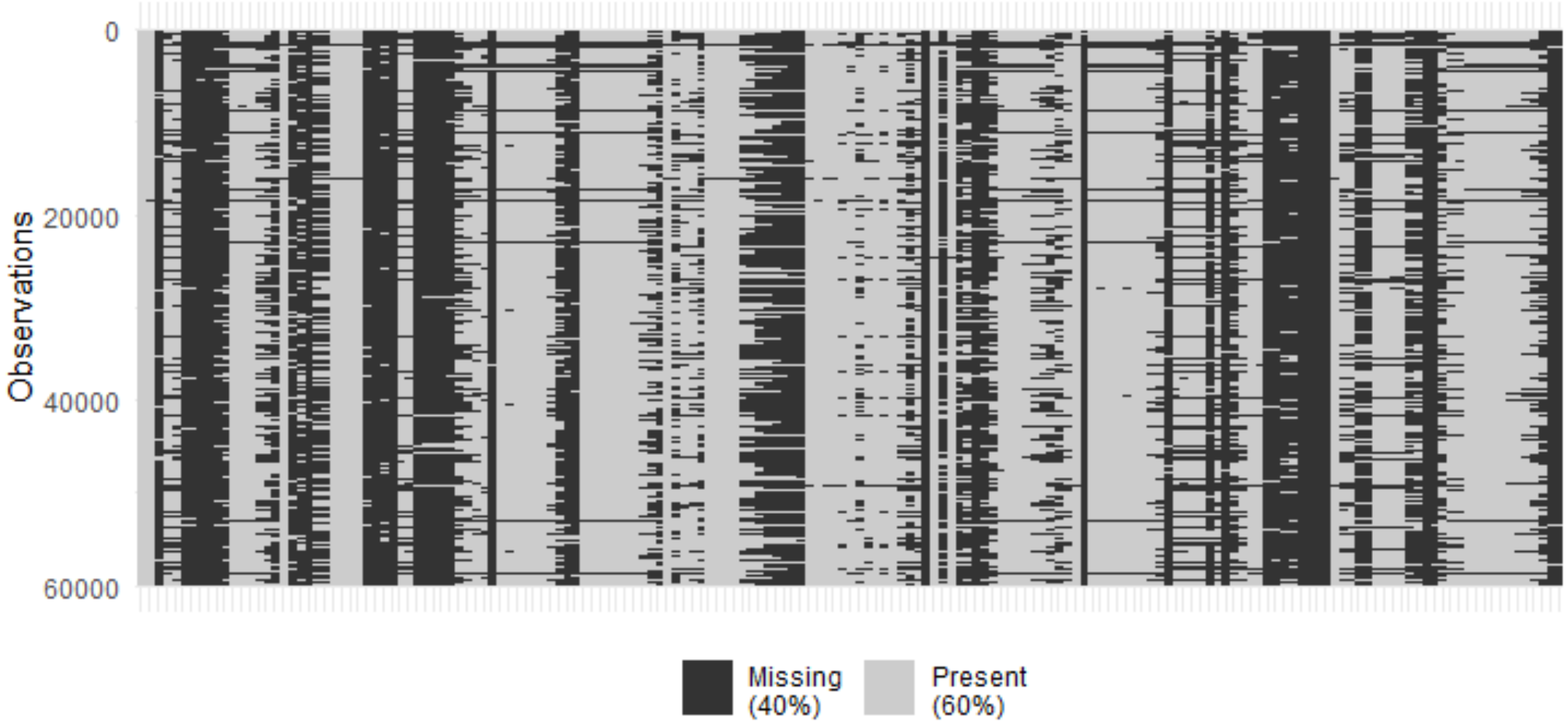
- ESTAVA ASSIM.



BIX TECNOLOGIA

AÇÕES DE LIMPEZA E TRANSFORMAÇÃO

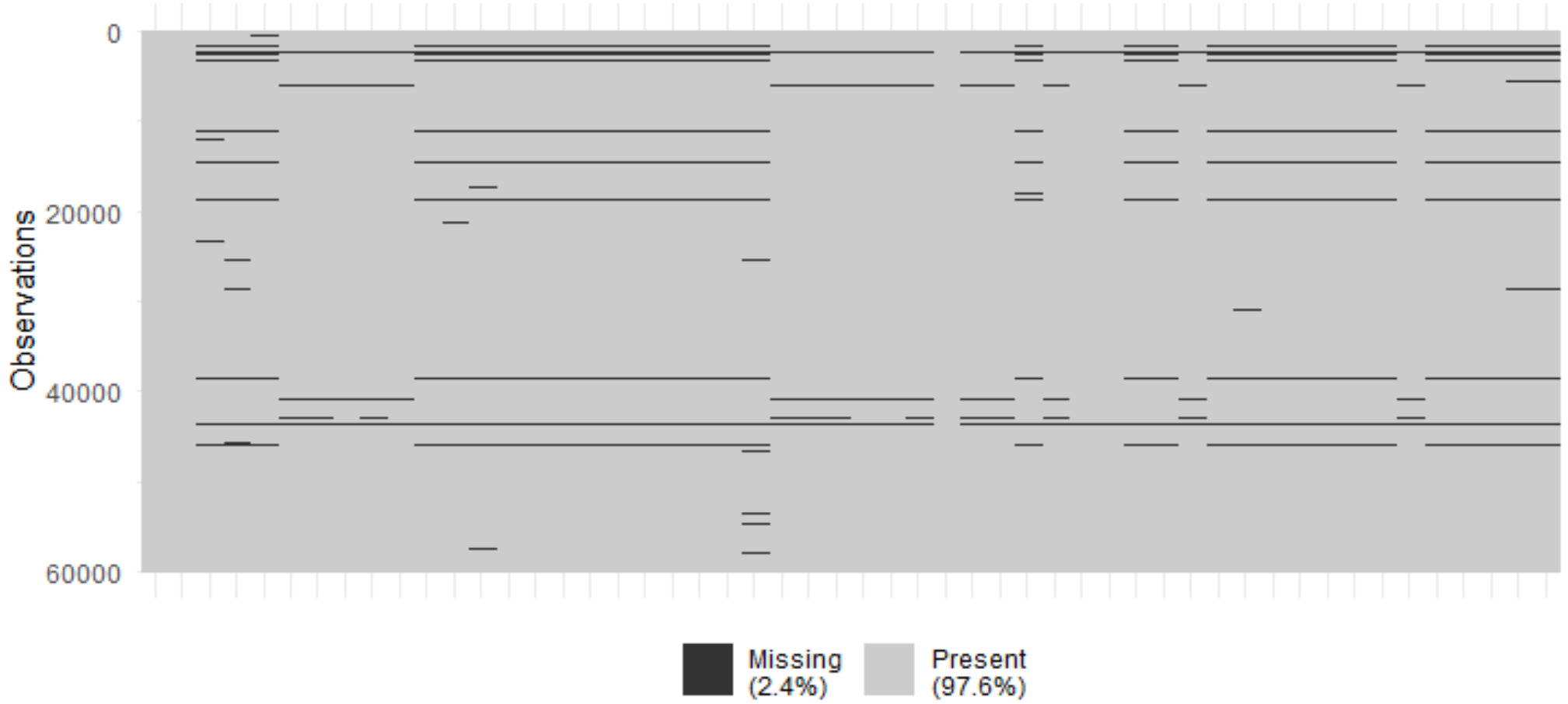
- PRECISOU FICAR ASSIM.



BIX TECNOLOGIA

AÇÕES DE LIMPEZA E TRANSFORMAÇÃO

- PARA ENTÃO FICAR ASSIM.



BIX TECNOLOGIA

AÇÕES DE LIMPEZA E TRANSFORMAÇÃO

- SUBISTITUIR VALORES “0” POR “NA”.
- CONTABILIZAR O TOTAL DE VALORES “NA” POR COLUNA.
- EXCLUIR COLUNAS COM MAIS DE 5% DE VALORES “NA”.
- DATA SETE RESULTANTE COM 52 COLUNAS, 2,4% DE VALORES “NA”.
- IMPUTAR A MEDIANA PARA OS DEMAIS DADOS “NA” RESTANTES.
- PADRONIZAR OS VALORES COM A FUNÇÃO “SCALE”.

BIX TECNOLOGIA

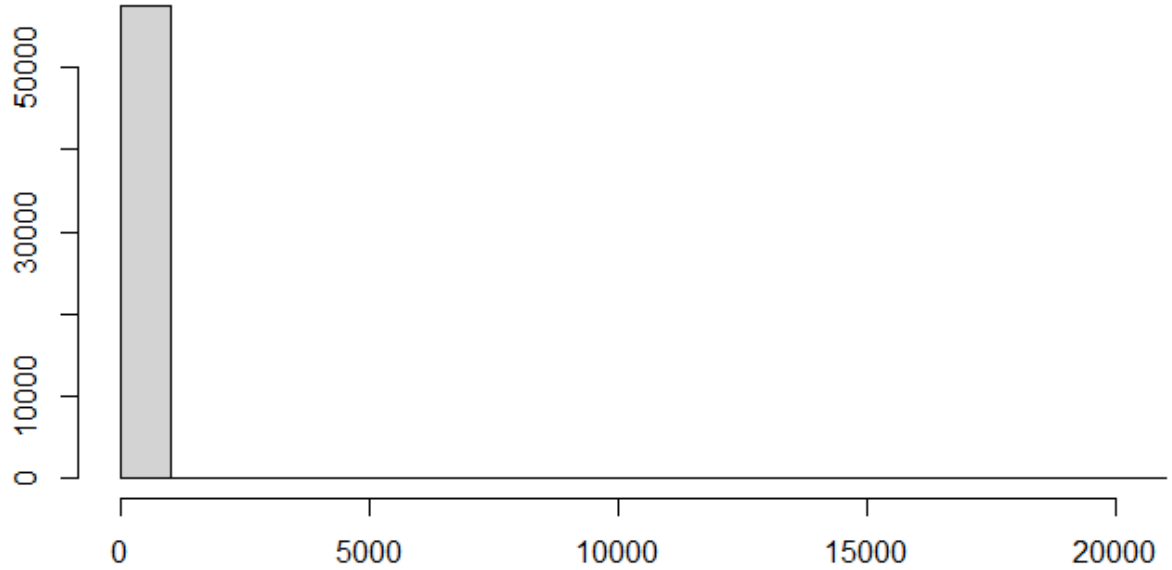
ANÁISE EXPLORATÓRIA

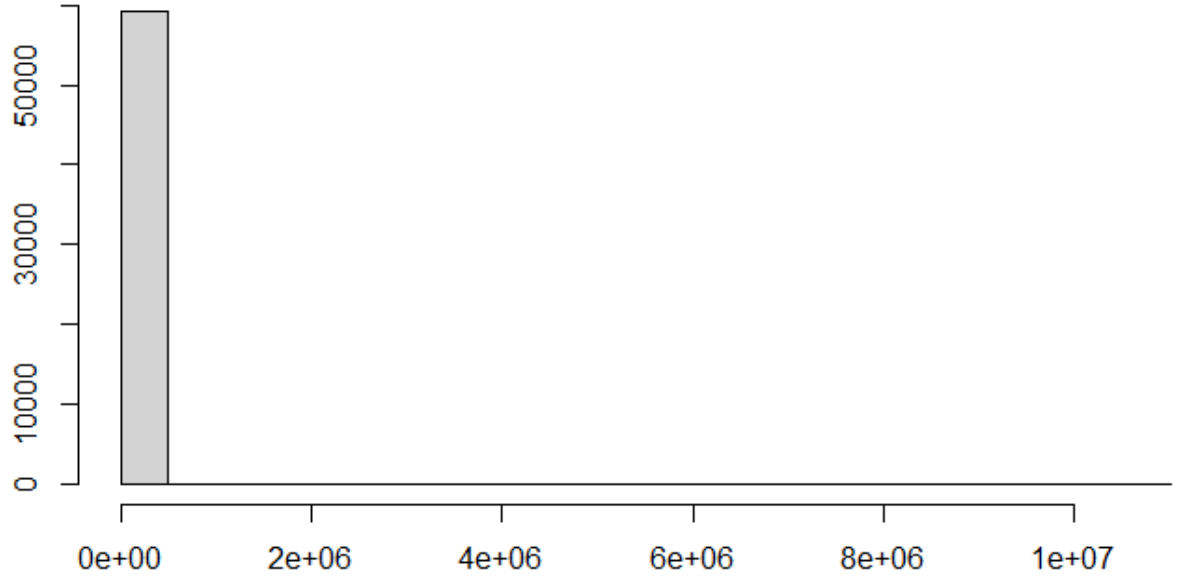
- As variáveis não apresentam correlação negativa;
- Os grupos az, bs e co apresentam FRACA correlação positiva;
- As demais apresentam FORTE correlação positiva;

BIX TECNOLOGIA

ANÁISE EXPLORATÓRIA

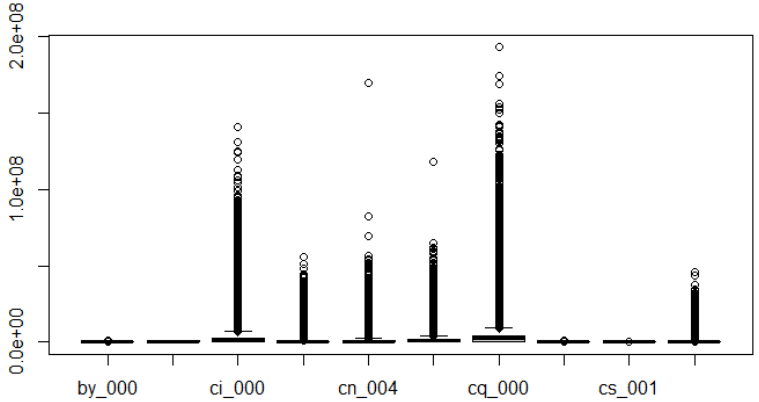
- As variáveis não apresentam distribuição normal;
- Muitas parecem sofrer interferência de valores discrepantes;

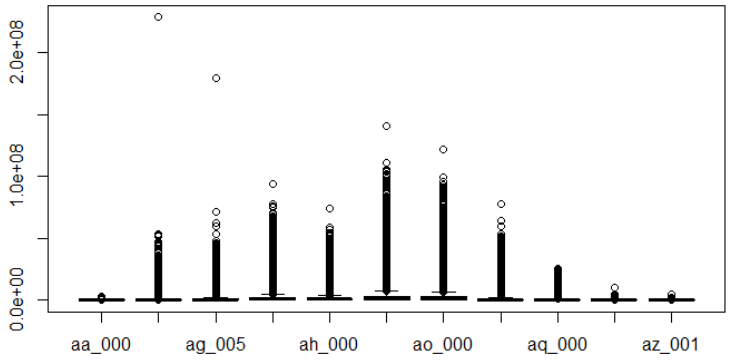


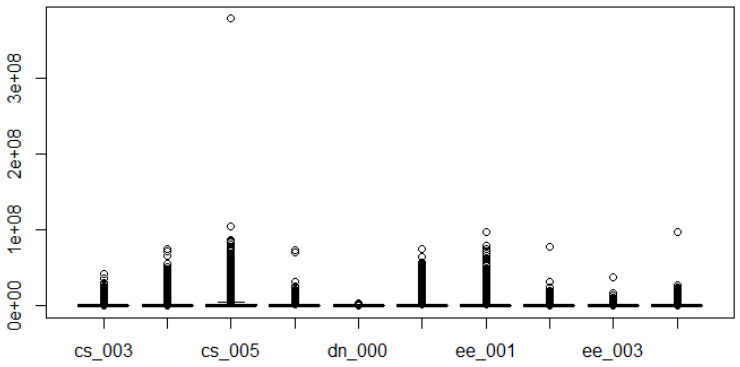


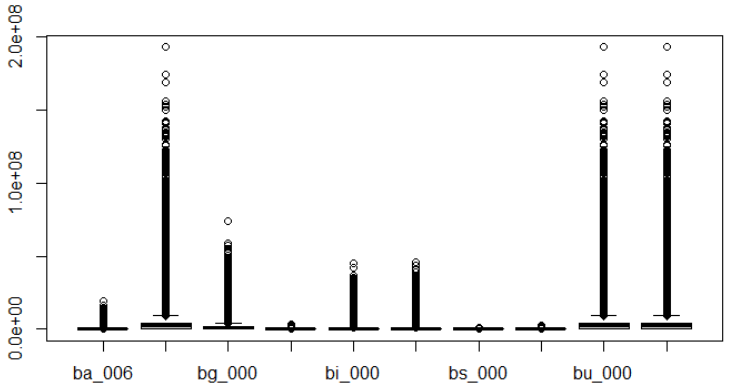
BIX TECNOLOGIA

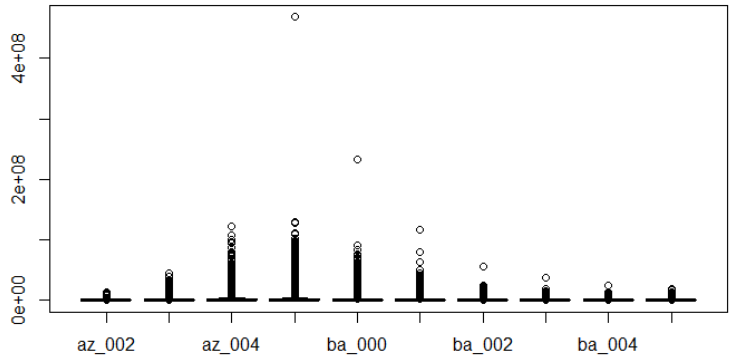
ANÁISE EXPLORATÓRIA











BIX TECNOLOGIA

ALGORITMOS APLICADOS

- REGRESSÃO LOGÍSTICA
- ARVORE DE DECISÃO
- (TENTATIVA) RENDON FOREST
- USANDO O MSAZURE:
  - Two-Class Decision Forest
  - Two-Class Support Vector Machine
  - Two-Class Logistic Regression (com PCA)
  - Two-Class Decision Jungle (com PCA)

BIX TECNOLOGIA

RESULTADOS ALCANÇADOS

- REGRESSÃO LOGÍSTICA
- NÍVEL DE ACURÁCIA – 97,2%
- MELHOR PARA PREVER O NÃO APRESENTA DEFEITO.

	não	sim
não	97,5% 15229	2,5% 396
sim	9,9% 37	90,1% 338

BIX TECNOLOGIA

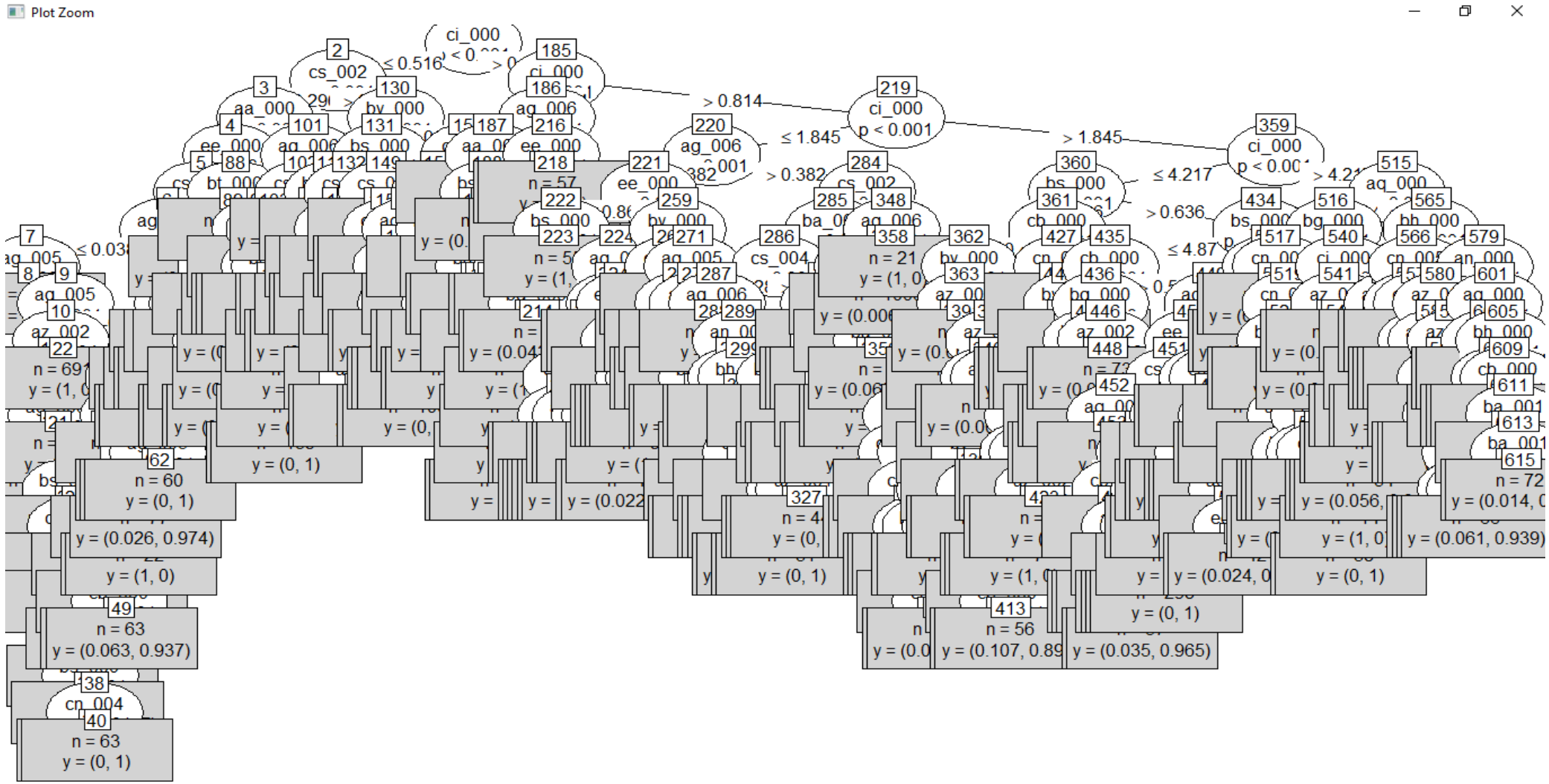
RESULTADOS ALCANÇADOS

- ÁRVORE DE DECISÃO
- NÍVEL DE ACURÁCIA – 97,4%
- MUITO RUIM PARA PREVER O NÃO APRESENTA DEFEITO.

	não	sim
não	99,4% 15298	0,6% 99
sim	54,2% 327	45,8% 276

BIX TECNOLOGIA

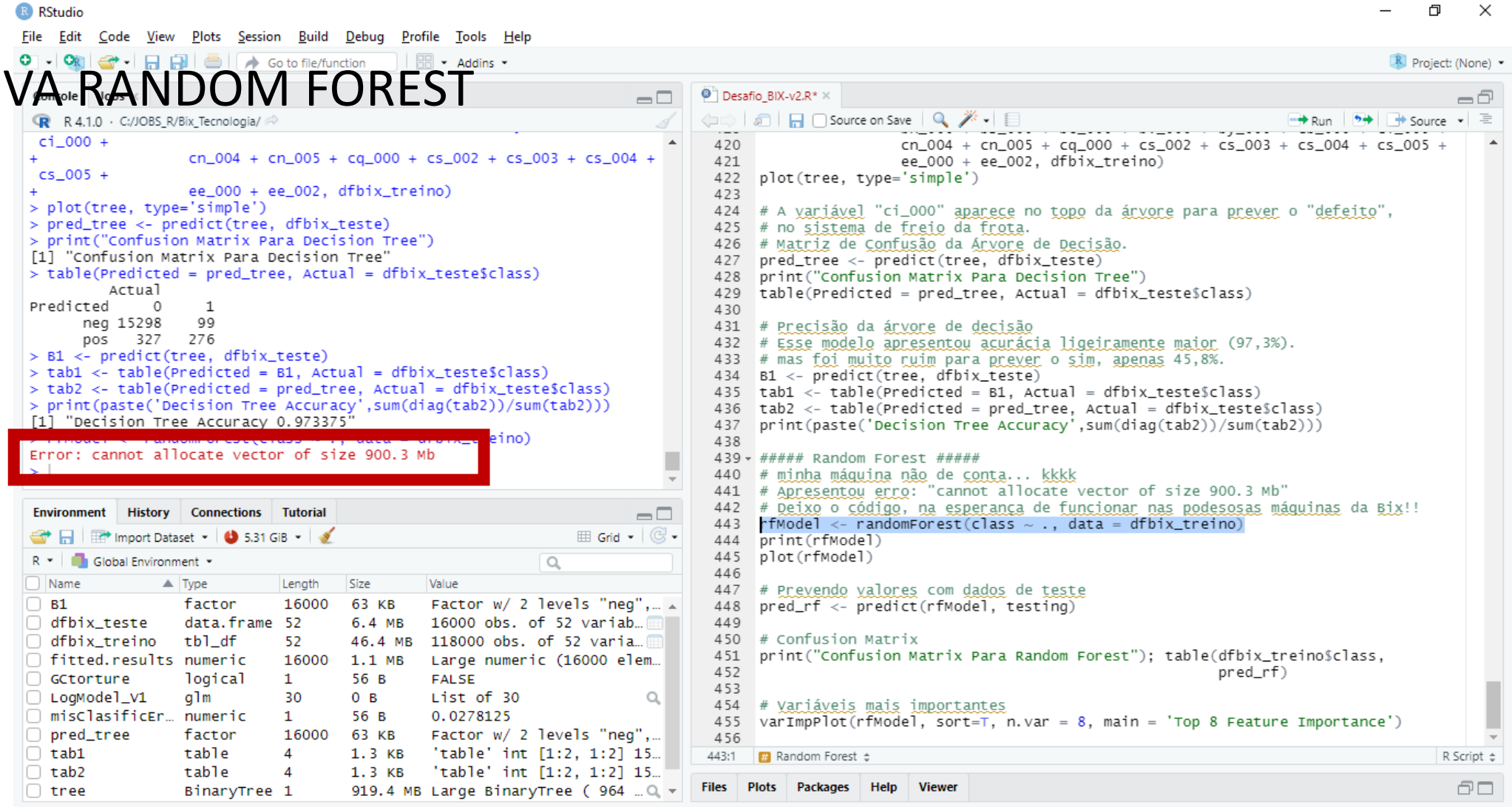
RESULTADOS ALCANÇADOS



BIX TECNOLOGIA

RESULTADOS ALCANÇADOS

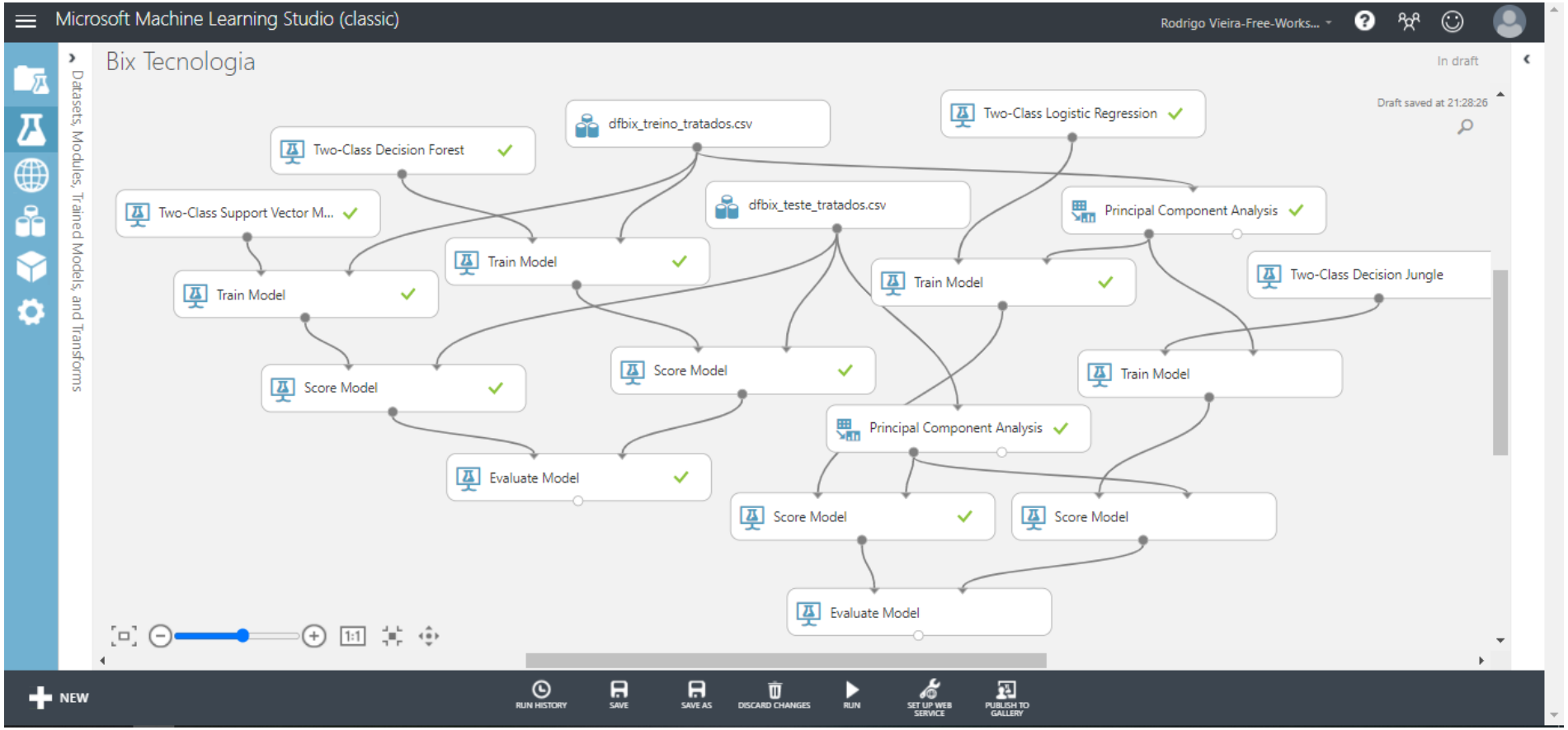
- TENTATIVA RANDOM FOREST



BIX TECNOLOGIA

RESULTADOS ALCANÇADOS

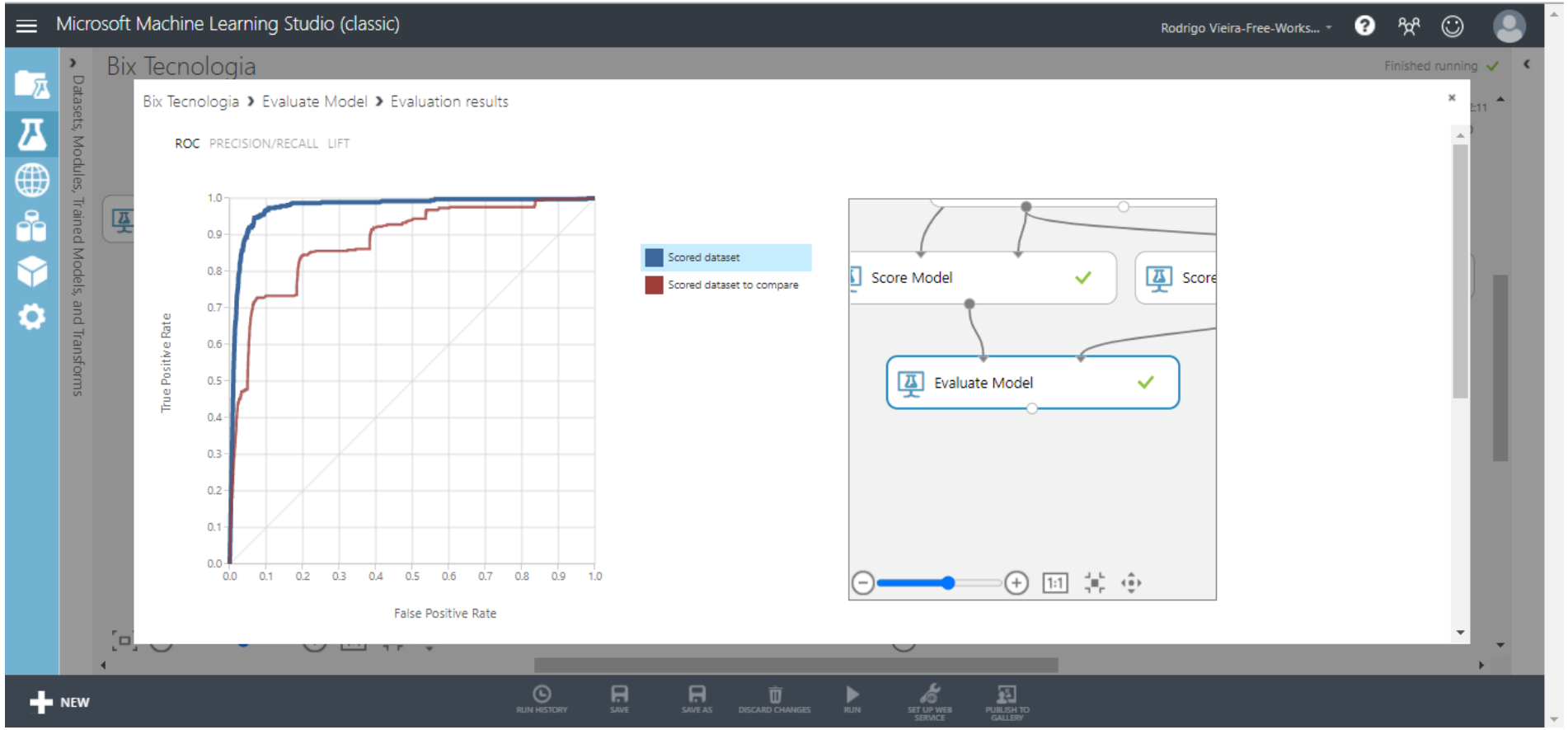
- USANDO O MSAZURE



BIX TECNOLOGIA

RESULTADOS ALCANÇADOS

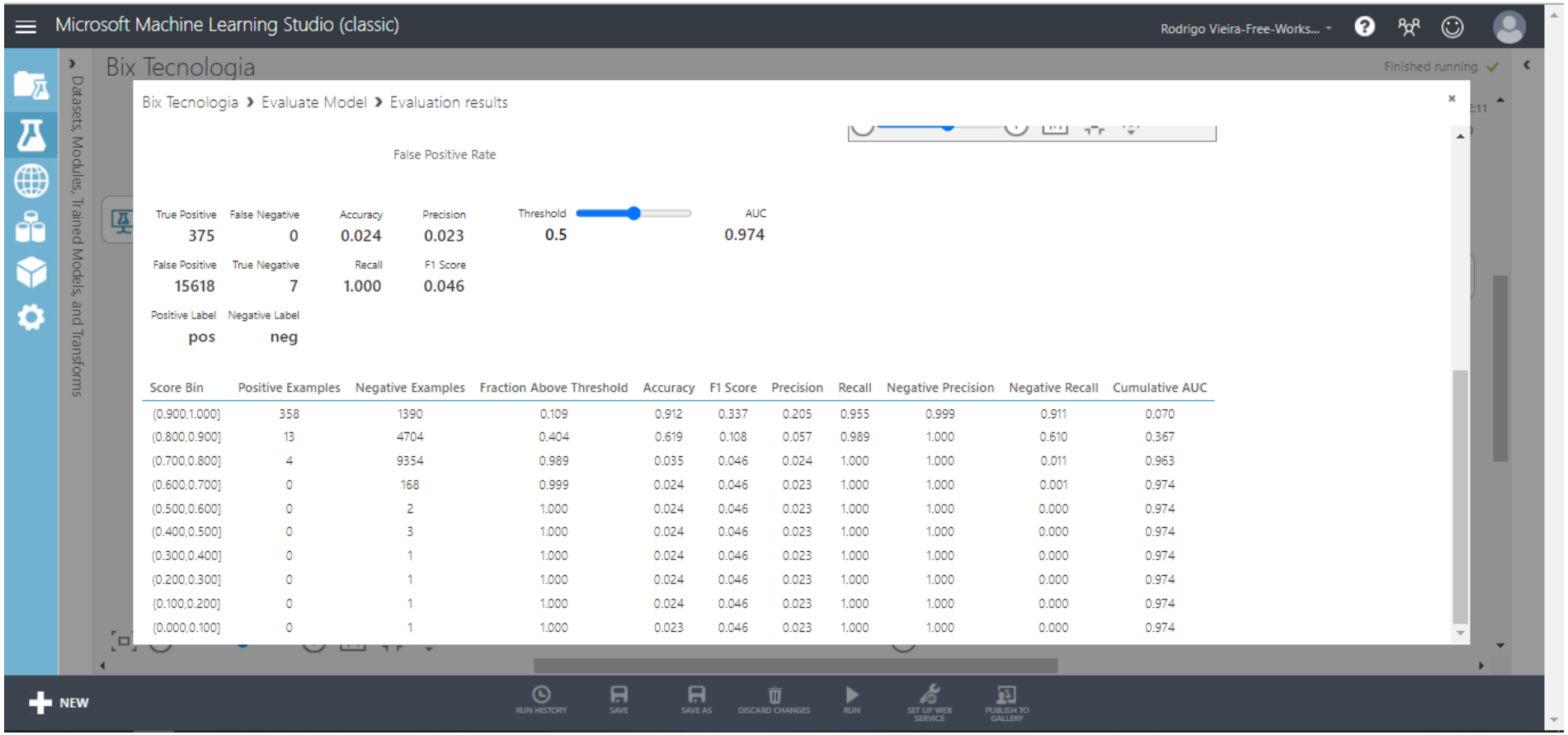
- USANDO O MSAZURE



BIX TECNOLOGIA

RESULTADOS ALCANÇADOS

- USANDO O MSAZURE

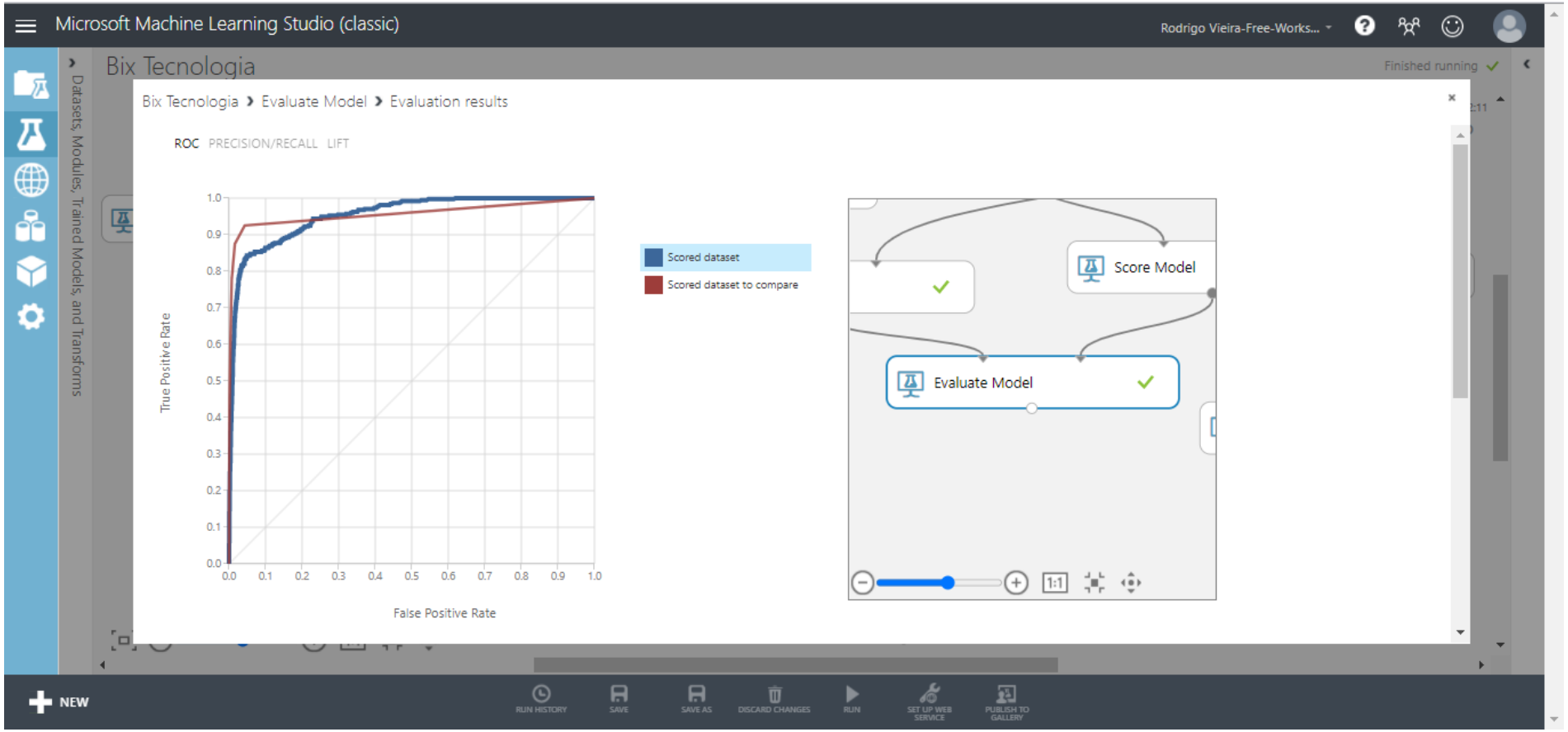


Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	358	1390	0.109	0.912	0.337	0.205	0.955	0.999	0.911	0.070
(0.800,0.900]	13	4704	0.404	0.619	0.108	0.057	0.989	1.000	0.610	0.367
(0.700,0.800]	4	9354	0.989	0.035	0.046	0.024	1.000	1.000	0.011	0.963
(0.600,0.700]	0	168	0.999	0.024	0.046	0.023	1.000	1.000	0.001	0.974
(0.500,0.600]	0	2	1.000	0.024	0.046	0.023	1.000	1.000	0.000	0.974
(0.400,0.500]	0	3	1.000	0.024	0.046	0.023	1.000	1.000	0.000	0.974
(0.300,0.400]	0	1	1.000	0.024	0.046	0.023	1.000	1.000	0.000	0.974
(0.200,0.300]	0	1	1.000	0.024	0.046	0.023	1.000	1.000	0.000	0.974
(0.100,0.200]	0	1	1.000	0.024	0.046	0.023	1.000	1.000	0.000	0.974
(0.000,0.100]	0	1	1.000	0.023	0.046	0.023	1.000	1.000	0.000	0.974

BIX TECNOLOGIA

RESULTADOS ALCANÇADOS

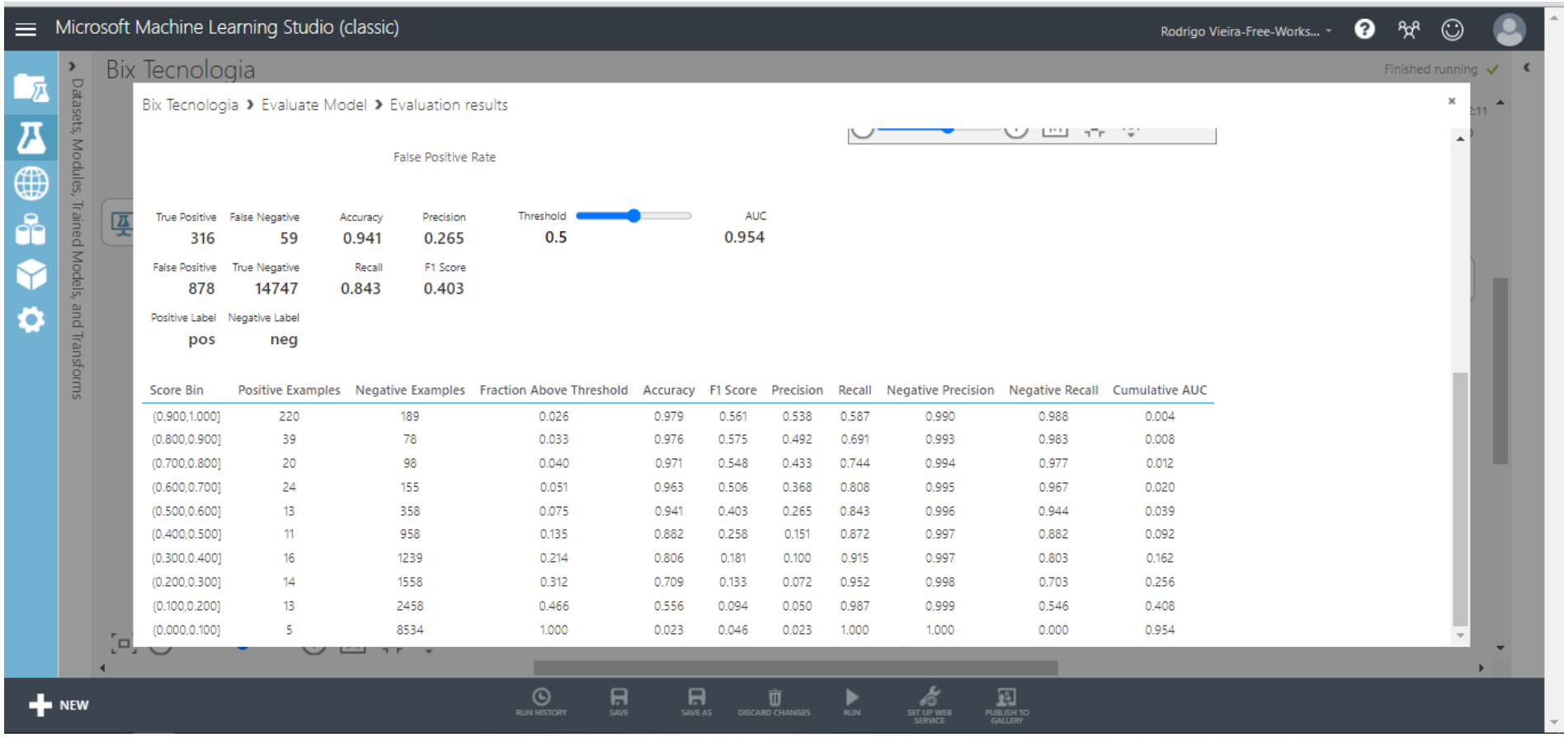
- USANDO O MSAZURE



BIX TECNOLOGIA

RESULTADOS ALCANÇADOS

- USANDO O MSAZURE



Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	220	189	0.026	0.979	0.561	0.538	0.587	0.990	0.988	0.004
(0.800,0.900]	39	78	0.033	0.976	0.575	0.492	0.691	0.993	0.983	0.008
(0.700,0.800]	20	98	0.040	0.971	0.548	0.433	0.744	0.994	0.977	0.012
(0.600,0.700]	24	155	0.051	0.963	0.506	0.368	0.808	0.995	0.967	0.020
(0.500,0.600]	13	358	0.075	0.941	0.403	0.265	0.843	0.996	0.944	0.039
(0.400,0.500]	11	958	0.135	0.882	0.258	0.151	0.872	0.997	0.882	0.092
(0.300,0.400]	16	1239	0.214	0.806	0.181	0.100	0.915	0.997	0.803	0.162
(0.200,0.300]	14	1558	0.312	0.709	0.133	0.072	0.952	0.998	0.703	0.256
(0.100,0.200]	13	2458	0.466	0.556	0.094	0.050	0.987	0.999	0.546	0.408
(0.000,0.100]	5	8534	1.000	0.023	0.046	0.023	1.000	1.000	0.000	0.954

BIX TECNOLOGIA

SUGESTÕES E RECOMENDAÇÕES

- REVER PROCESSO DE COLETA DE DADOS PARA REDUZIR PERDAS DE REGISTROS.