

Projeto Interdisciplinar Big Data + ML

Titanic Problem aplicado à última eleição brasileira

- Pós Graduação em Ciência de Dados - 2022.2
- IFSP Campinas
- Profa. Bianca Pedrosa - bpedrosa@ifsp.edu.br
Prof. Samuel Martins (Samuka) - @hisamuka
- Outubro de 2022
- aluno: Swift Yaguchi - CP301665X



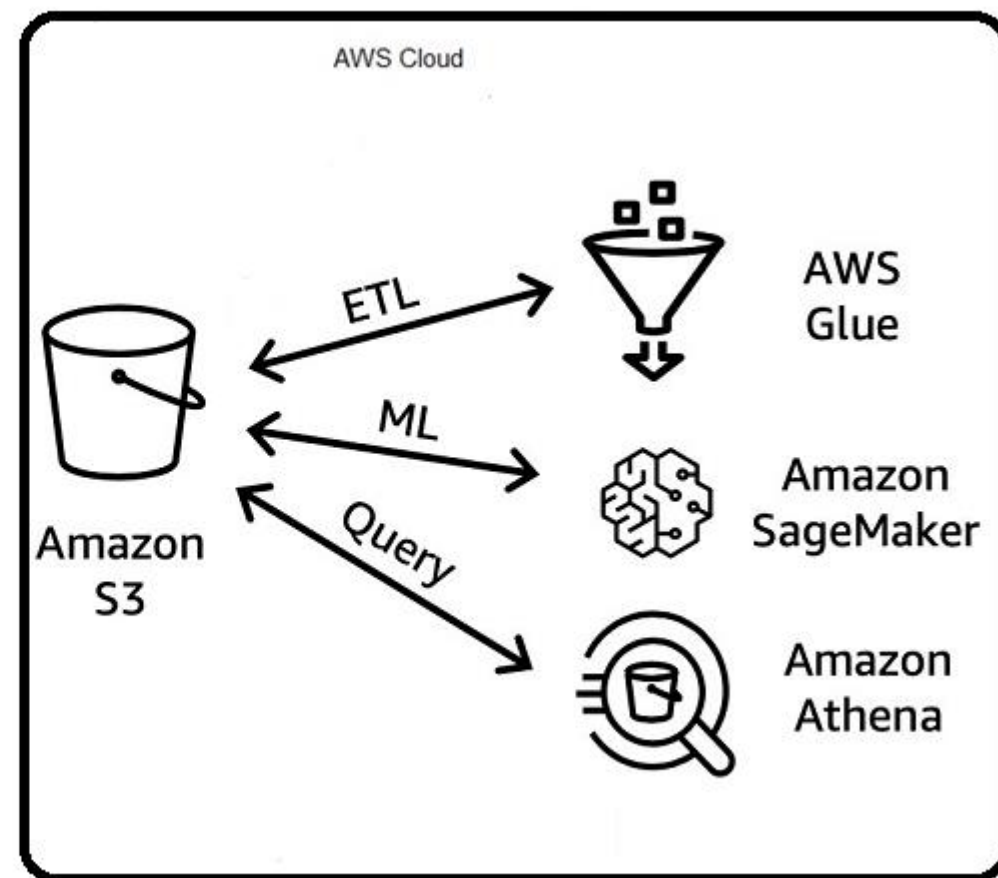
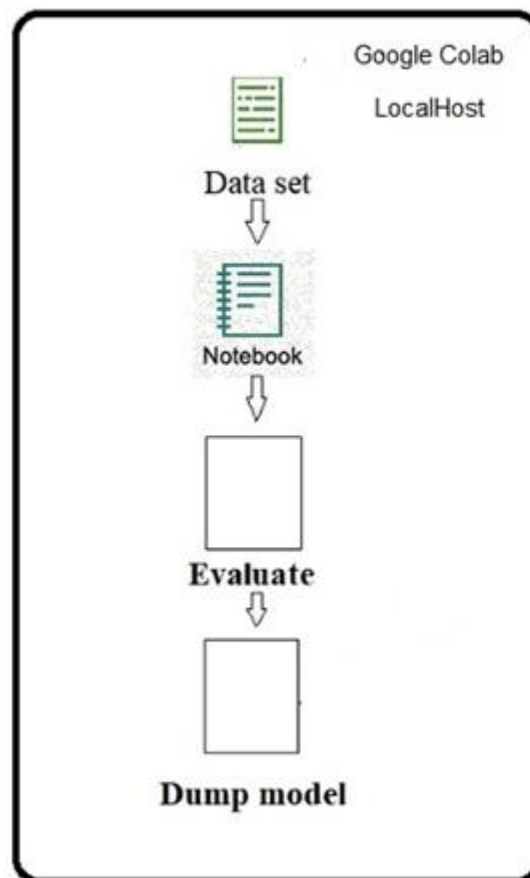
Proposta do Projeto - "Frame the Problem"

- Inspiração :
 - "Titanic Problem" na plataforma do Kaggle
- Titanic survival prediction:
 - dados dos passageiros do navio
>> *características dos sobreviventes do naufrágio.*
- Neste projeto:
 - dados de candidatos da eleição brasileira de 2022
>> *características dos candidatos eleitos.*
- Objetivo: Desenvolver um modelo ML para previsão estimada de candidatos eleitos para legislativo federal e estadual



Arquitetura e Fluxo de Trabalho

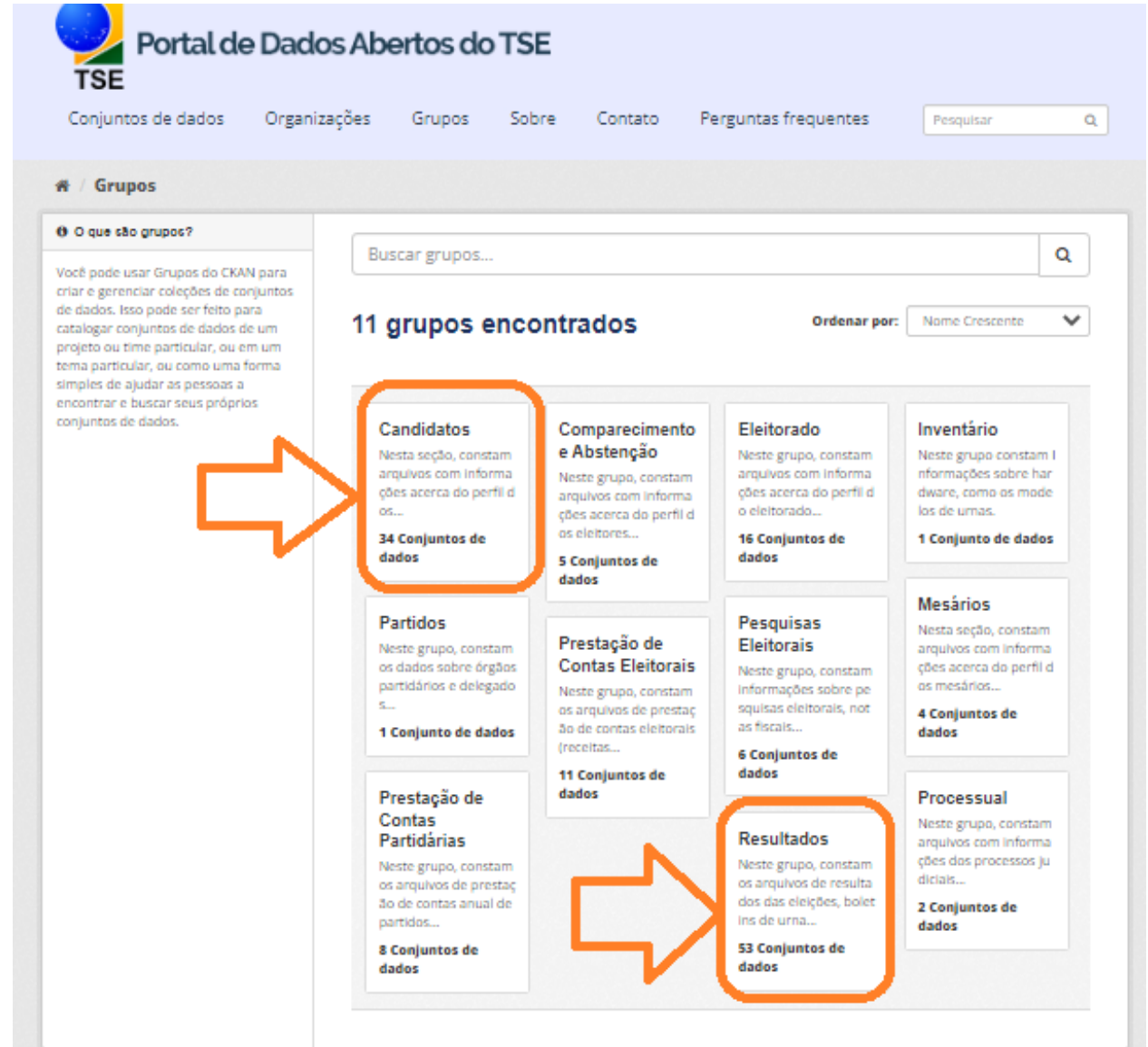
- Desenvolvimento inicial do notebook Jupyterlab:
 - *ambiente do Google Colab*
 - *ambiente local do meu computador.*
- Carregado no AWS Cloud para finalização
 - *Fluxo de trabalho ML no Sagemaker*
 - *Dados carregados em bucket S3*
 - *ETL de dados com AWS Glue*
 - *Queries com Athena*



Get Data

Perfil dos candidatos
(5.986KB zip)

Resultados de votação por
município e por zona eleitoral
(830.299KB zip)



The screenshot shows the 'Portal de Dados Abertos do TSE' website. The header includes the TSE logo and navigation links: 'Conjuntos de dados', 'Organizações', 'Grupos', 'Sobre', 'Contato', and 'Perguntas frequentes'. A search bar is located on the right. The main content area is titled 'Grupos' and displays '11 grupos encontrados'. A sidebar on the left explains the CKAN Groups feature. The main grid lists various data groups, with 'Candidatos' and 'Resultados' highlighted by orange arrows.

Grupo	Descrição	Conjuntos de dados
Candidatos	Nesta seção, constam arquivos com informações acerca do perfil dos...	34 Conjuntos de dados
Comparecimento e Abstenção	Neste grupo, constam arquivos com informações acerca do perfil dos eleitores...	5 Conjuntos de dados
Eleitorado	Neste grupo, constam arquivos com informações acerca do perfil do eleitorado...	16 Conjuntos de dados
Inventário	Neste grupo constam informações sobre hardware, como os modelos de urnas.	1 Conjunto de dados
Mesários	Nesta seção, constam arquivos com informações acerca do perfil dos mesários...	4 Conjuntos de dados
Processual	Neste grupo, constam arquivos com informações dos processos judiciais...	2 Conjuntos de dados
Pesquisas Eleitorais	Neste grupo, constam informações sobre pesquisas eleitorais, notas fiscais...	6 Conjuntos de dados
Prestação de Contas Eleitorais	Neste grupo, constam os arquivos de prestação de contas eleitorais (receitas...	11 Conjuntos de dados
Partidos	Neste grupo, constam os dados sobre órgãos partidários e delegados...	1 Conjunto de dados
Prestação de Contas Partidárias	Neste grupo, constam os arquivos de prestação de contas anual de partidos...	8 Conjuntos de dados
Resultados	Neste grupo, constam os arquivos de resultados das eleições, boletins de urna...	53 Conjuntos de dados

Get Data

Perfil dos candidatos (5.986KB zip)

Resultados de votação por município e por zona eleitoral (830.299KB zip)



Limpeza e Pré-processamento

- 13165 – train
- 3292 - test

RangeIndex: 13165 entries, 0 to 13164

Data columns (total 12 columns):

#	Column	Non-Null Count	Dtype
0	CD_CARGO	13165 non-null	int64
1	SQ_CANDIDATO	13165 non-null	int64
2	NR_IDADE_DATA_POSSE_x	13165 non-null	float64
3	CD_GENERO	13165 non-null	int64
4	CD_GRAU_INSTRUCAO	13165 non-null	int64
5	CD_ESTADO_CIVIL	13165 non-null	int64
6	CD_COR_RACA	13165 non-null	int64
7	CD_OCUPACAO	13165 non-null	int64
8	ST_REELEICAO_x	13165 non-null	int64
9	VR_BEM_CANDIDATO_x	13165 non-null	float64
10	QT_VOTOS_NOMINAIS_x	13165 non-null	int64
11	DS_SIT_TOT_TURN0	13165 non-null	int64

dtypes: float64(2), int64(10)

memory usage: 1.2 MB

RangeIndex: 3292 entries, 0 to 3291

Data columns (total 12 columns):

#	Column	Non-Null Count	Dtype
0	CD_CARGO	3292 non-null	int64
1	SQ_CANDIDATO	3292 non-null	int64
2	NR_IDADE_DATA_POSSE_x	3292 non-null	float64
3	CD_GENERO	3292 non-null	int64
4	CD_GRAU_INSTRUCAO	3292 non-null	int64
5	CD_ESTADO_CIVIL	3292 non-null	int64
6	CD_COR_RACA	3292 non-null	int64
7	CD_OCUPACAO	3292 non-null	int64
8	ST_REELEICAO_x	3292 non-null	int64
9	VR_BEM_CANDIDATO_x	3292 non-null	float64
10	QT_VOTOS_NOMINAIS_x	3292 non-null	int64
11	DS_SIT_TOT_TURN0	3292 non-null	int64

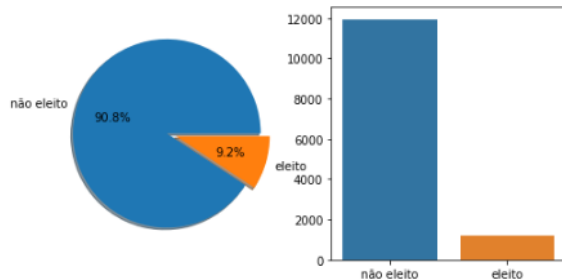
dtypes: float64(2), int64(10)

memory usage: 308.8 KB

Get Data

Feature	Tipo	Descrição
CD_CARGO	Int64	Deputado Federal, Estadual, Senador
SQ_CANDIDATO	Int64	Nº de identificação única do candidato na base de dados
NR_IDADE_DATA_POSSE	float64	Idade do candidato no ano da posse, se eleito
CD_GENERO	Int64	Masculino, feminino, não declarado
CD_GRAU_INSTRUCAO	int64	Analfabeto, le e escreve, fundamental, médio, superior
CD_ESTADO_CIVIL	Int64	Solteiro, casado, viúvo, separado judicialmente, divorciado
CD_COR_RACA	Int64	Branca, preta, parda, amarela, indígena, não informado
CD_OCUPACAO	int64	Profissões diversas, inclusive vereador, deputado
ST_REELEICAO	Int64	0 não é candidato a reeleição, 1 é candidato a reeleição
VR_BEM_CANDIDATO	Float64	Valor declarado de patrimônio do candidato
QT_VOTOS_NOMINAIS	Int64	Nº de votos obtidos no 1º turno da eleição 2022
DS_SIT_TOT_TURNO (target feature)	Int64	1 – eleito, eleito por QE, eleito por média 0 – suplente, não eleito

Exploratory Data Analysis (EDA):



- Apenas 9,2% dos candidatos são eleitos

- Maior taxa de eleitos:

- *Candidatos à reeleição*

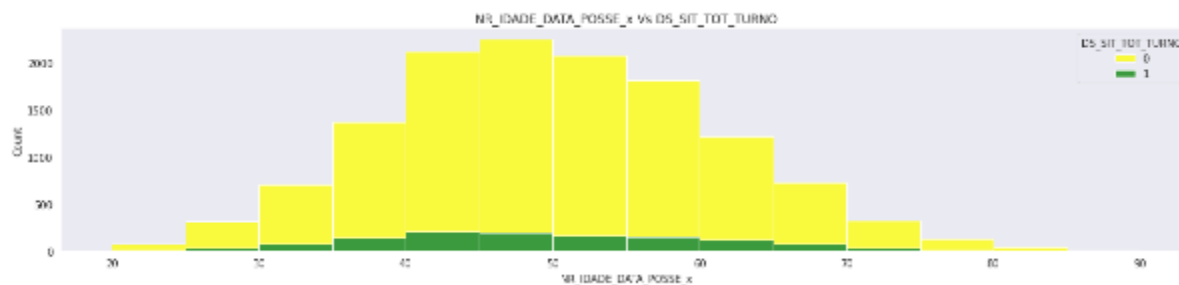
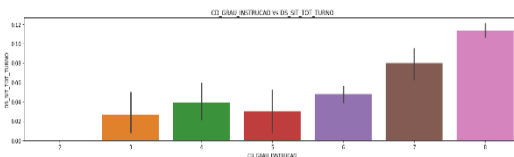
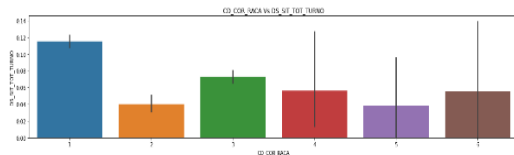
- *Faixa de idade entre 40 e 60 anos, e são os mais eleitos*

- *Branco*

- *Casados*

- *Grau de instrução superior*

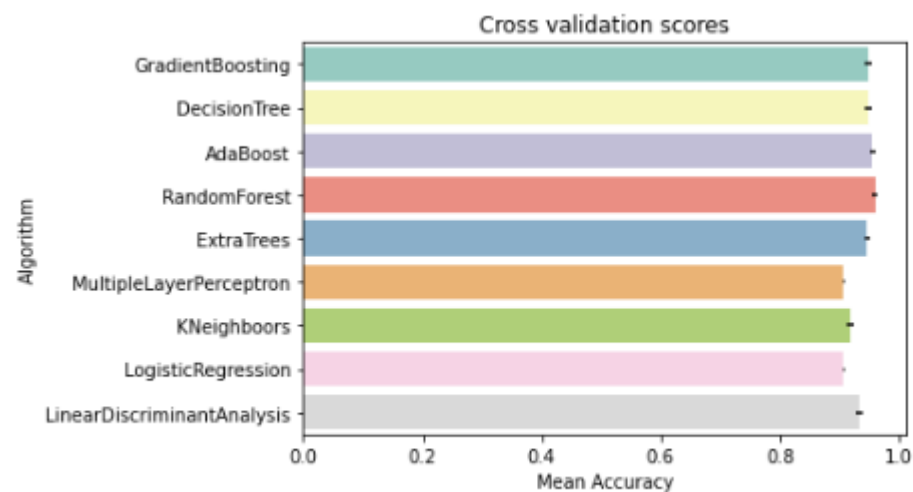
- *Homens*



Treinamento Machine Learning

Baseline

Cross Validation



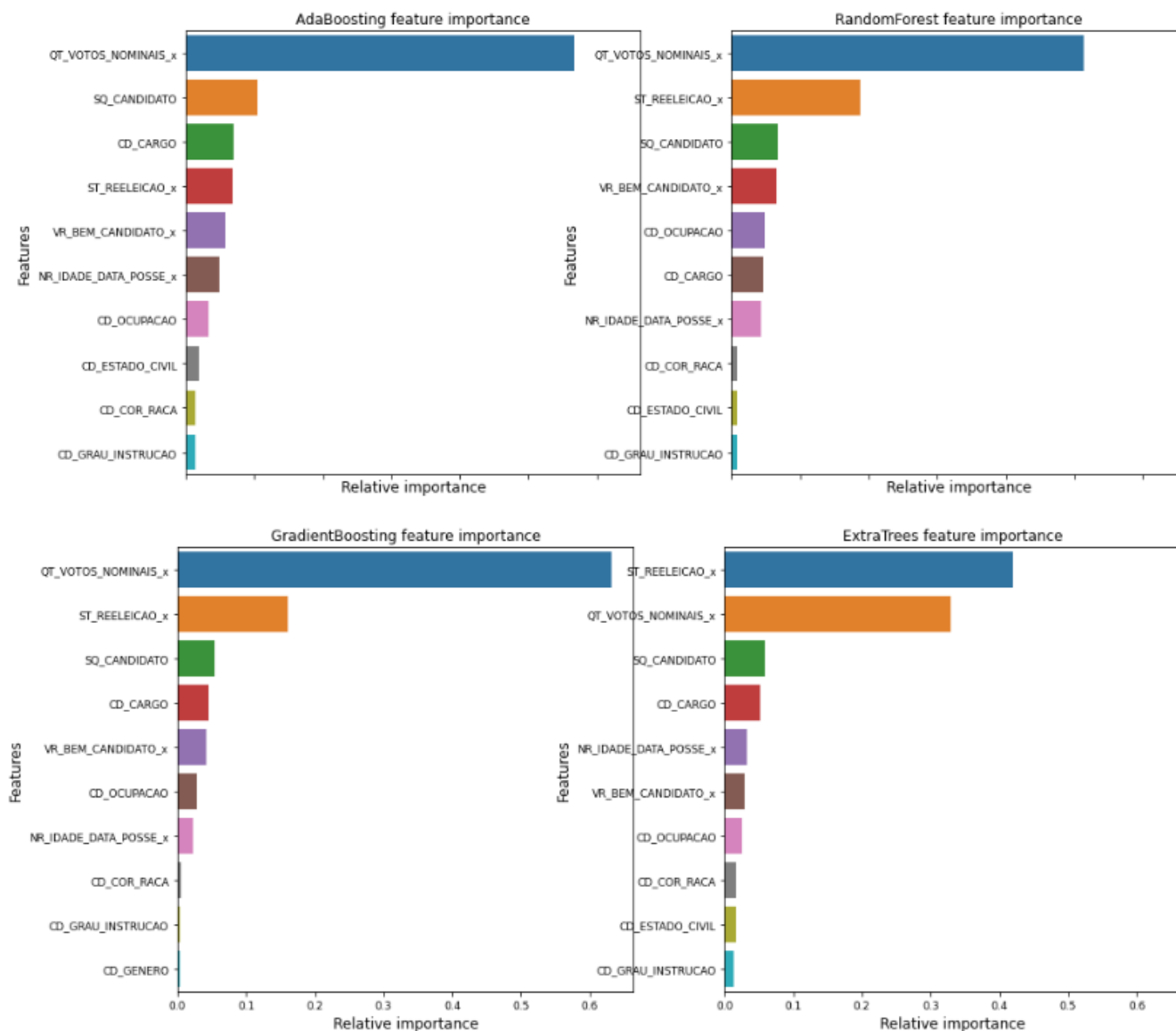
[15]:

	CrossValMeans	CrossValerrors	Algorithm
3	0.959742	0.004876	RandomForest
2	0.955260	0.005178	AdaBoost
1	0.948273	0.006755	DecisionTree
0	0.948273	0.007205	GradientBoosting
4	0.946373	0.004164	ExtraTrees
8	0.933840	0.006913	LinearDiscriminantAnalysis
6	0.918039	0.005036	KNeighbors
7	0.907634	0.000344	LogisticRegression
5	0.907330	0.001054	MultipleLayerPerceptron

Resultado Emsemble Modeling:

0.9577764277035237

Análise das variáveis

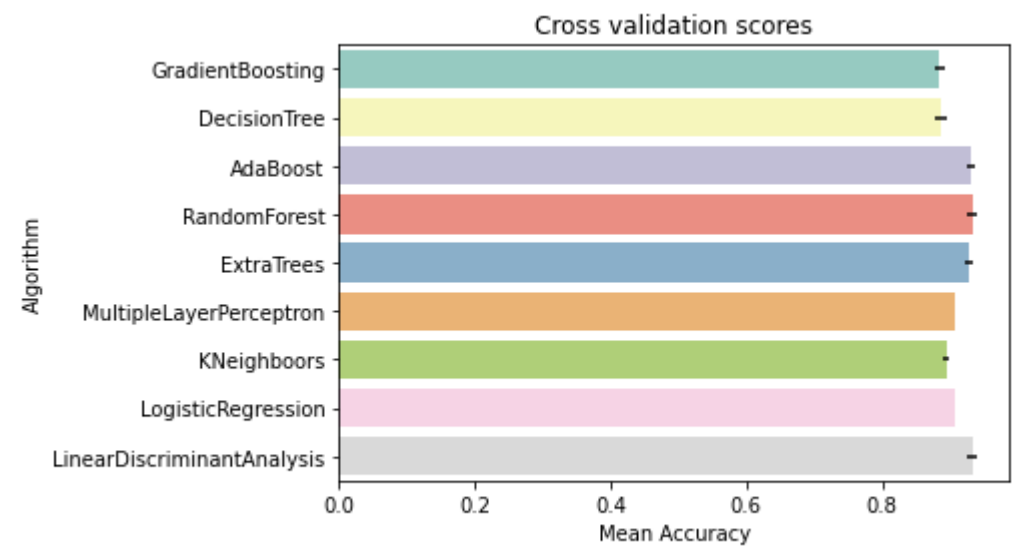


AdaBoosting,
Random Forest e
GradientBoosting

➤ Quantidade de votos
nominais tem maior
importância

Treinamento Machine Learning

Retirando variável :
número de votos recebidos

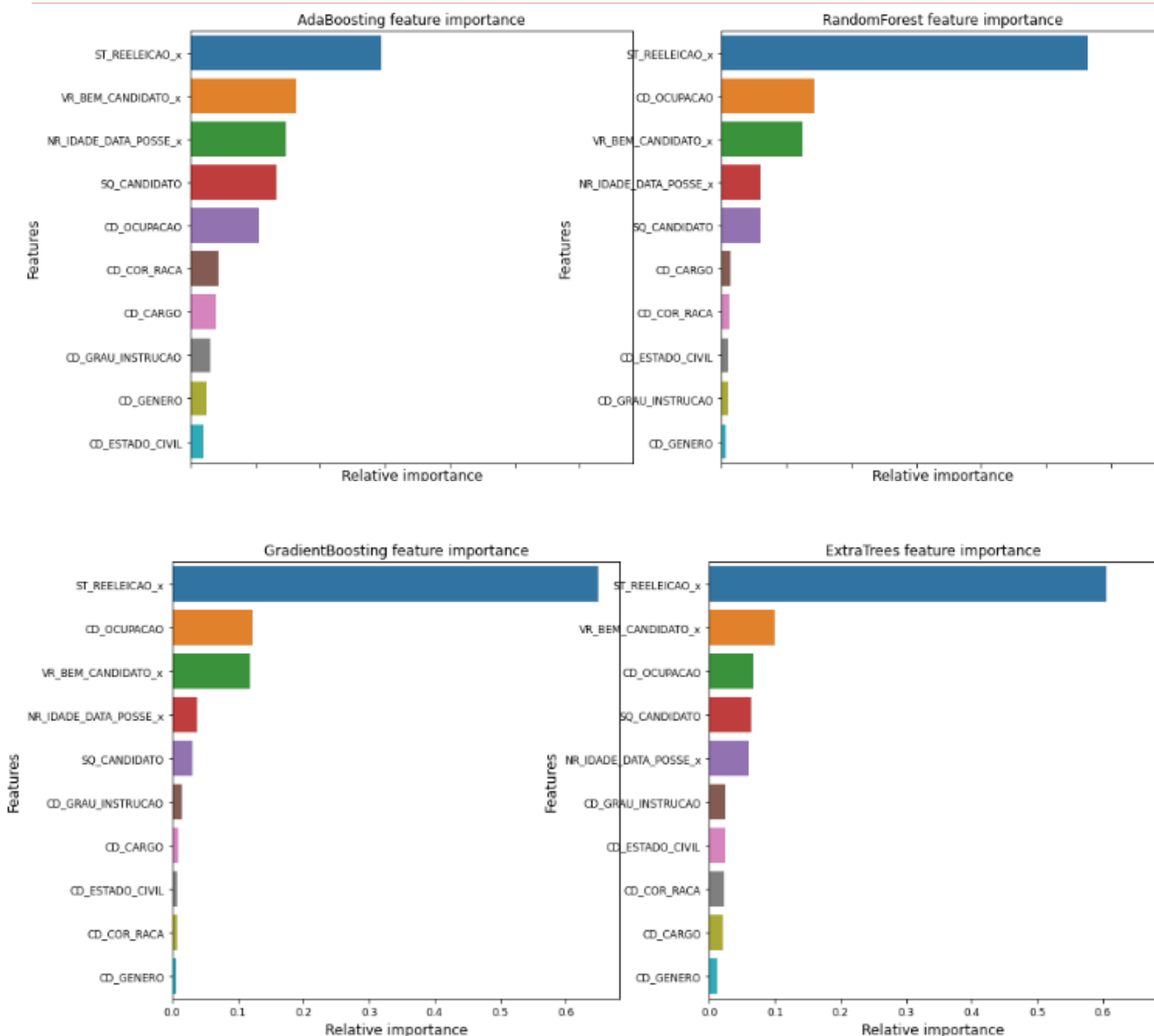


Resultado Emsemble Modeling:

0.925577156743621

	CrossValMeans	CrossValerrors	Algorithm
8	0.933992	0.007103	LinearDiscriminantAnalysis
3	0.933385	0.006939	RandomForest
2	0.931865	0.005090	AdaBoost
4	0.928523	0.005194	ExtraTrees
7	0.907634	0.000344	LogisticRegression
5	0.907406	0.000963	MultipleLayerPerceptron
6	0.894493	0.003329	KNeighbors
1	0.886897	0.008041	DecisionTree
0	0.885378	0.007638	GradientBoosting

Análise das variáveis



'ST_REELEICAO_x' tem maior importância para todos

Em seguida:

AdaBoosting e Extra Trees:

- VR_BEM_CANDIDATO com importância relativa maior que CD_OCUPACAO,

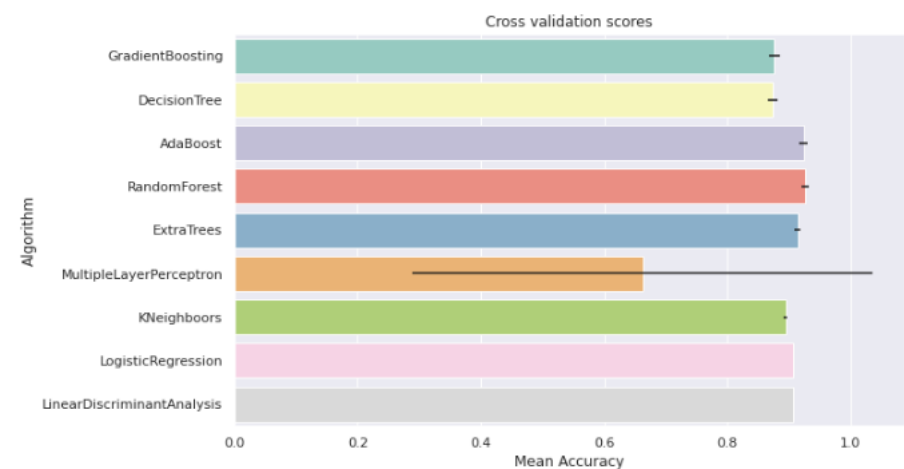
Random Forest e Gradient Boosting

- O contrário.

Treinamento Machine Learning

Estudo eliminando as variáveis:

- QT_VOTOS_NOMINAIS_x
- ST_REELEICAO_x



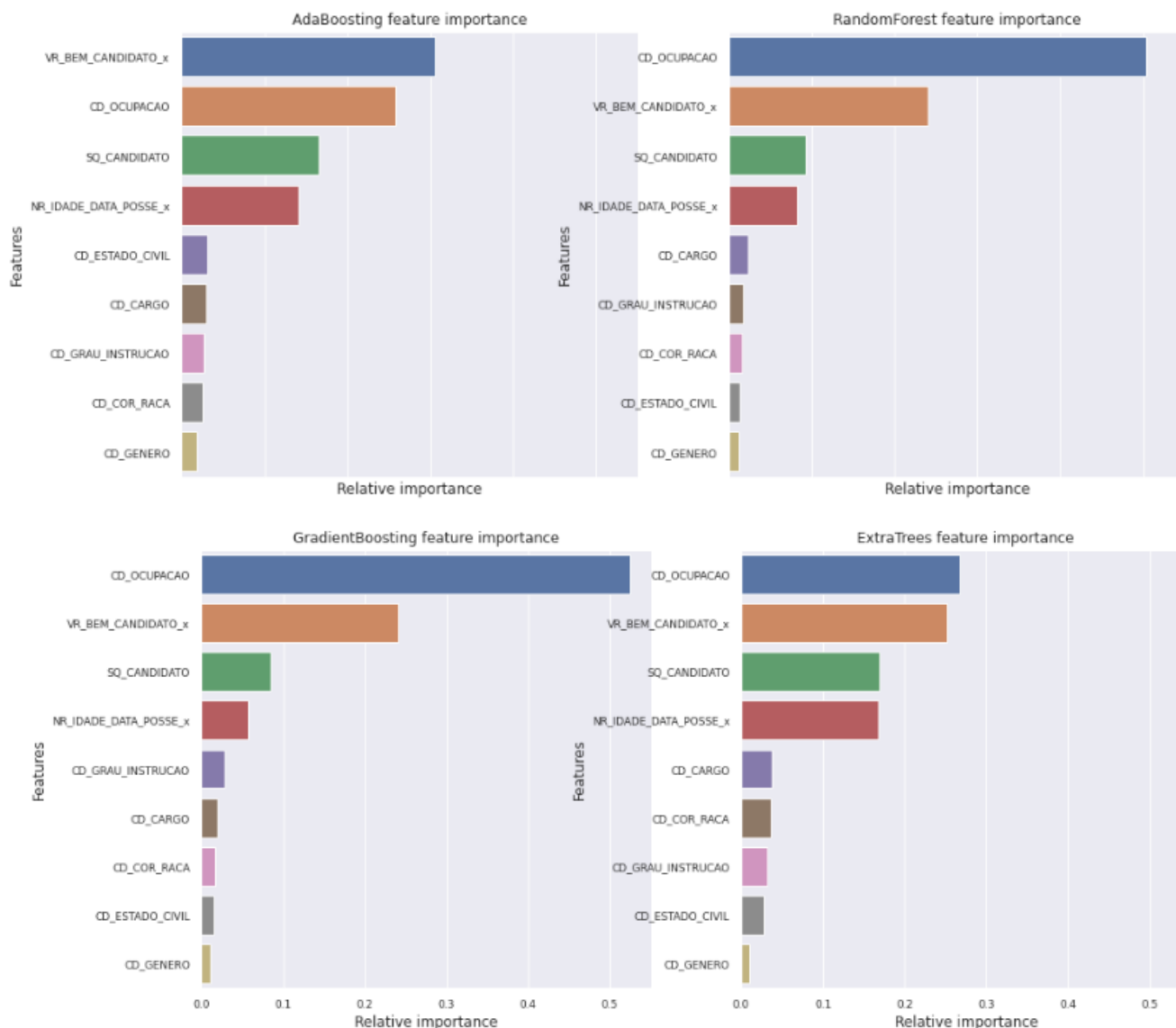
Resultado Emsemble Modeling:

0.9170716889428918

[18]:

	CrossValMeans	CrossValerrors	Algorithm
3	0.926244	0.005791	RandomForest
2	0.923890	0.006302	AdaBoost
4	0.914319	0.004322	ExtraTrees
7	0.907634	0.000344	LogisticRegression
8	0.907254	0.000706	LinearDiscriminantAnalysis
6	0.894493	0.003329	KNeighbors
0	0.876795	0.008635	GradientBoosting
1	0.873681	0.007695	DecisionTree
5	0.663001	0.373544	MultipleLayerPerceptron

Análise das variáveis



AdaBoosting :

➤ VR_BEM_CANDIDATO

com importância relativa maior que
CD_OCUPACAO,

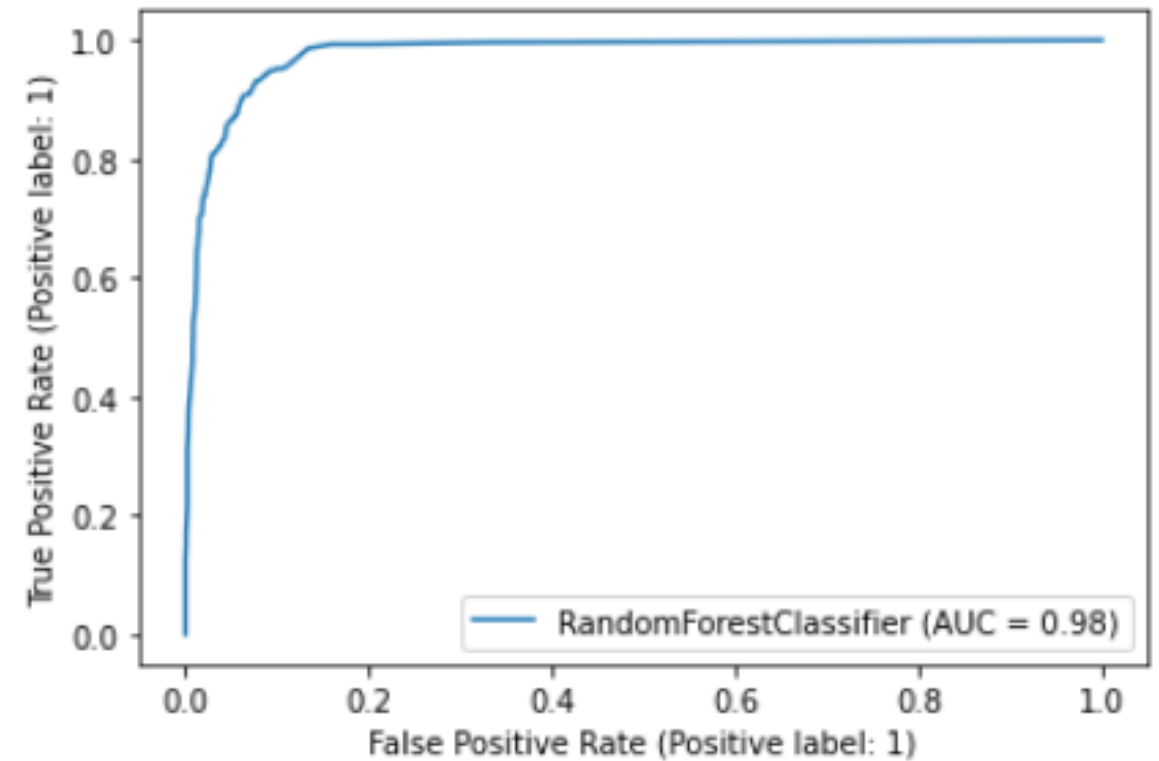
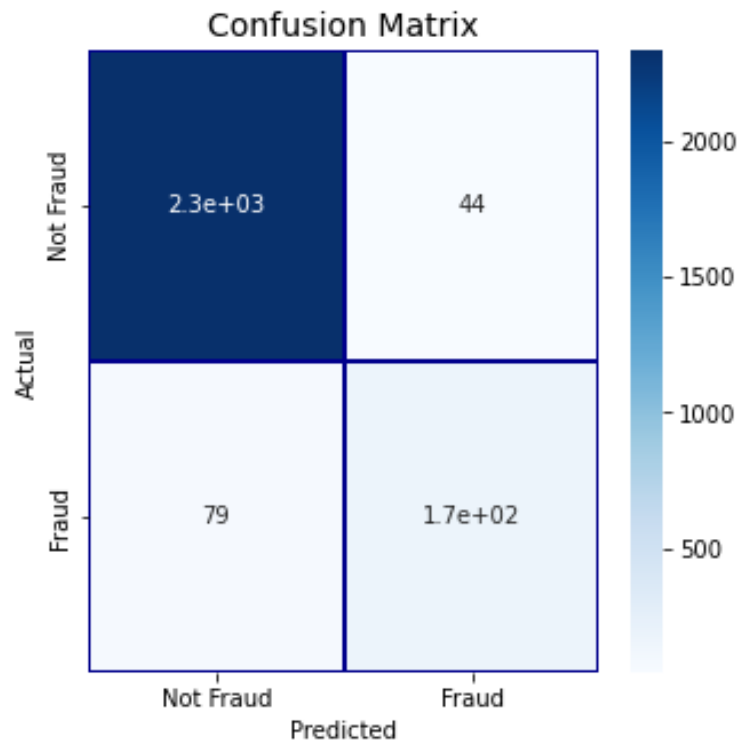
Random Forest,
Extra Trees e
Gradient Boosting :

➤ O contrário.

Conclusão

Estudos	Candidatos Eleitos	Score de Acurácia
Baseline	242	0.9577764277035237
Eliminado variável - nº de votos recebidos	170	0.925577156743621
Eliminando as variáveis: - qt_votos_nominais - st_reeleicao	94	0.9170716889428918

Estudo adicional – ROC AOC

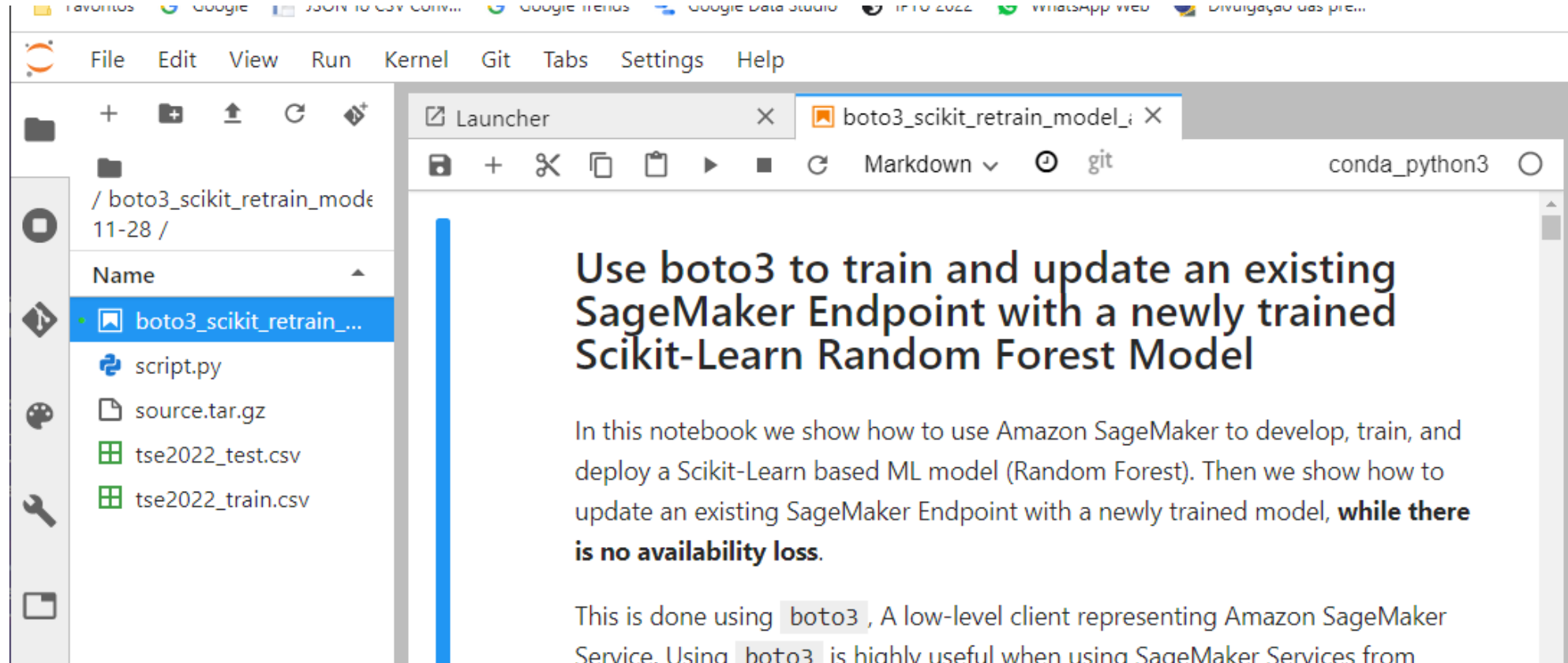


Resultado ROC-AOC

Com ROC-AUC obtivemos melhoria na acurácia:

0.956798592978226

Estudo adicional – AWS Sagemaker



Estudo adicional – AWS Sagemaker

8.2 Preparação dos dados

```
#####  
# Reading the data (from AWS S3)  
#####  
session = boto3.Session()  
s3 = session.client('s3')  
  
data_bucket_name = "syaguchi-aws-ifspcps-bucket-tse2022"  
#obj_list = s3_client.list_objects(Bucket=data_bucket_name)  
obj_list = s3.list_objects(Bucket=data_bucket_name)  
file = []  
for contents in obj_list["Contents"]:  
    file.append(contents["Key"])  
print(file)
```

```
['Proj_Int_2022.2/test/tse2022_test.csv', 'Proj_Int_2022.2/train/tse2022_train  
lidade.csv', 'bem_candidato_2022_BRASIL.csv', 'consulta_cand_2022_BRASIL.csv'  
1-2022-11-28-03-55-02/output/model.tar.gz', 'sagemaker/sklearn-tse2022/tse202  
022_train.csv', 'scikitlearn-tse-2022-train-from-boto3/source.tar.gz', 'sciki  
z', 'test_df.csv/test_df.csv', 'train_df.csv/train_df.csv', 'votacao_candidat
```

```
file_data = 'votacao_candidato_munzona_2022_BRASIL_pre_merge13.csv'  
response = s3.get_object(Bucket=data_bucket_name, Key=file_data)  
response_body = response["Body"].read()  
data = pd.read_csv(io.BytesIO(response_body), header=0, delimiter=";", low_m
```

```
target = 'ELEITO'  
predictors = ['CD_CARGO', 'NR_IDADE_DATA_POSSE', 'CD_GENERO', 'CD_GRAU_INSTR  
'CD_COR_RACA', 'CD_OCUPACAO', 'CD_REELEICAO', 'VR_BEM_CANDIDATO']
```

8.7 Deploy do modelo

```
[16]: endpoint_name = "sklearn-endpoint-" + datetime.datetime.now().strftime("%Y-%m-%d-%H-%M-%S")  
  
create_endpoint_response = client.create_endpoint(  
    EndpointName=endpoint_name,  
    EndpointConfigName=endpoint_config_1_name,  
)  
  
create_endpoint_response
```

```
t[16]: {'EndpointArn': 'arn:aws:sagemaker:us-east-1:887118459548:endpoint/sklearn-endpoint-2022-11-28-04-44-  
'ResponseMetadata': {'RequestId': '546d0c7a-5db0-4bda-9fdd-1efb641d1461',  
'HTTPStatusCode': 200,  
'HTTPHeaders': {'x-amzn-requestid': '546d0c7a-5db0-4bda-9fdd-1efb641d1461',  
'content-type': 'application/x-amz-json-1.1',  
'content-length': '104',  
'date': 'Mon, 28 Nov 2022 04:44:01 GMT'},  
'RetryAttempts': 0}}
```

```
[17]: describe_endpoint_response = client.describe_endpoint(EndpointName=endpoint_name)  
  
while describe_endpoint_response["EndpointStatus"] == "Creating":  
    describe_endpoint_response = client.describe_endpoint(EndpointName=endpoint_name)  
    print(describe_endpoint_response["EndpointStatus"])  
    time.sleep(15)  
  
describe_endpoint_response
```

```
Creating  
Creating  
Creating  
Creating  
Creating  
Creating  
Creating  
Creating  
InService
```

```
t[17]: {'EndpointName': 'sklearn-endpoint-2022-11-28-04-44-01',  
'EndpointArn': 'arn:aws:sagemaker:us-east-1:887118459548:endpoint/sklearn-endpoint-2022-11-28-04-44-  
'ResponseMetadata': {'RequestId': '546d0c7a-5db0-4bda-9fdd-1efb641d1461',  
'HTTPStatusCode': 200,  
'HTTPHeaders': {'x-amzn-requestid': '546d0c7a-5db0-4bda-9fdd-1efb641d1461',  
'content-type': 'application/x-amz-json-1.1',  
'content-length': '104',  
'date': 'Mon, 28 Nov 2022 04:44:01 GMT'},  
'RetryAttempts': 0}}
```

Estudo adicional – AWS Sagemaker

8.8 Clean-up

```
In [19]: sm_boto3.delete_endpoint(EndpointName=endpoint_name)
```

```
Out[19]: {'ResponseMetadata': {'RequestId': '6fcd17d7-e4e7-42e4-9ba3-61aac25dc5d7',  
    'HTTPStatusCode': 200,  
    'HTTPHeaders': {'x-amzn-requestid': '6fcd17d7-e4e7-42e4-9ba3-61aac25dc5d7',  
    'content-type': 'application/x-amz-json-1.1',  
    'content-length': '0',  
    'date': 'Mon, 28 Nov 2022 04:47:18 GMT'},  
    'RetryAttempts': 0}}
```

```
In [ ]:
```