PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Artur Ribeiro Filho

PROPOSTA DE MODELO PARA DETECÇÃO DE FRAUDES EM COMPRAS DA ADMINISTRAÇÃO PÚBLICA FEDERAL

Belo Horizonte
2021

Artur Ribeiro Filho

PROPOSTA DE MODELO PARA DETECÇÃO DE FRAUDES EM COMPRAS DA ADMINISTRAÇÃO PÚBLICA FEDERAL

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2021

SUMÁRIO

1. Introdução

1.1. Contextualização

O Poder Público no Brasil efetua todos os anos compras que envolvem valores consideráveis. De acordo com o site http://paineldecompras.economia.gov.br (acesso em 04 fev. 2021), que exibe um painel com os principais indicadores de compras, entre janeiro de 2017 até janeiro de 2020 o valor total pago em compras foi da ordem de aproximadamente 53 bilhões de reais com mais de 160 mil fornecedores participantes e com mais de 72 mil contratos celebrados.

Muitas empresas se aproveitam este ambiente para de alguma forma tentarem burlar as regras e obter vantagens indevidas. Os órgãos de controle efetuam auditorias constantes com intuito de evidenciar indicativos de fraudes e aplicar sanções às empresas que comprovadamente obtiveram algum ganho indevido.

É um trabalho de suma importância, dado que envolve dinheiro dos contribuintes de impostos e que devem ser retornados em benefícios à população.

Assim, este trabalho tem como objetivo contribuir para demonstrar de alguma forma a possibilidade de detectar possíveis fraudes e auxiliar no processo de controle.

1.2. O problema proposto

Com a quantidade gigantesca de dados que são produzidos atualmente, e nos processos de compras governamentais não é diferente, é imperativo o desenvolvimento de novas formas para agilizar a detecção de fraudes com objetivo de minimizar as perdas de dinheiro do erário. Assim a proposta deste trabalho é obter e analisar dados históricos e desenvolver um modelo de classificação que ajude a prever possíveis indicativos de irregularidades.

Pode-se utilizar da técnica dos 5-Ws para ajudar a entender o problema e a solução proposta, que consiste em responder às seguintes questões:

- Why? Auxiliar os órgãos de controle a monitorar de forma mais efetiva e ágil possíveis irregularidades que causem perdas consideráveis ao erário.

- Who? Os dados a serem analisados foram disponibilizados em bases abertas (públicas) por órgãos e sites do governo federal, como o Portal da Transparência.

- What? Identificar possíveis irregularidades por meio da análise de dados históricos de compras efetuadas pela Administração Pública Federal direta.

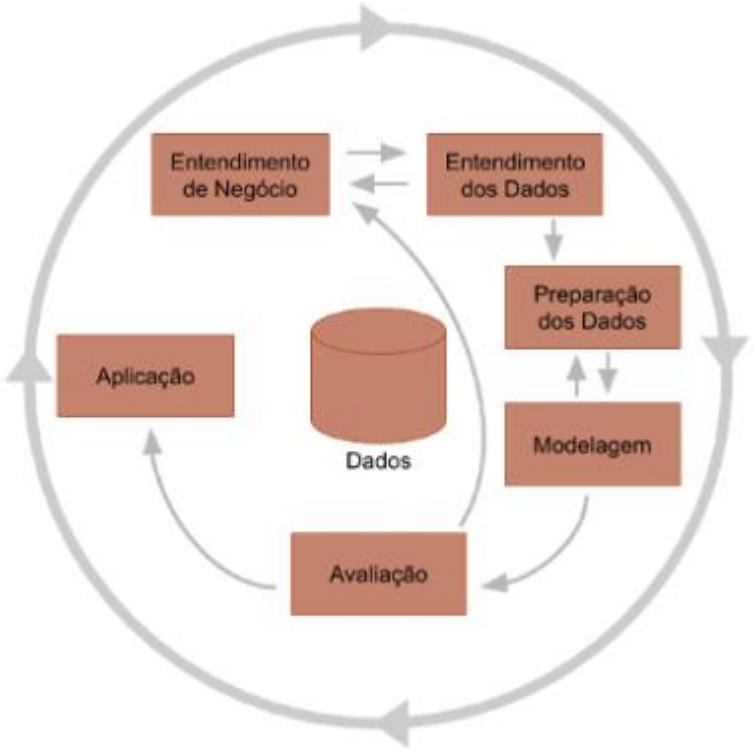
- Where? Serão analisados dados de contratos em nível nacional e apenas aqueles celebrados com pessoas jurídicas.

- When? O período considerado para a análise foi de janeiro de 2020 a novembro de 2020.

Para a execução de projetos de diversos tipos nas organizações, geralmente aplica-se alguma metodologia consagrada no mercado, tais como o PMBOK® do PMI – Project Management Institute ou Agile Project Management com SCRUM. Em projetos de Data Science também há uma metodologia consagrada que pode ser adotada para um melhor desempenho. Trata-se do Modelo de Referência CRISP-DM que é uma metodologia amplamente utilizada para execução de projetos de Data Science.

Este modelo apresenta uma visão geral do ciclo de vida de um projeto de Data Science contendo 6 fases que não necessariamente deverão ser seguidas rigorosamente mas que ajudam a diminuir o risco de fracasso pois serve como um guia de melhores práticas.

A seguir uma ilustração do ciclo de vida da metodologia CRISP-DM:



Fases do modelo CRISP-DM

1. Entendimento do Negócio: consiste em definir os objetivos do projeto sempre levando-se em conta os objetivos de negócio. Afinal é para isso que servem os projetos, para ajudar as organizações a resolverem problemas de negócio;

2. Entendimento dos Dados: consiste desde a captura dos dados até a identificação de problemas de qualidade nos dados.

3. Preparação dos Dados: consiste em preparar os dados para a modelagem. Aqui cria-se um conjunto de dados obtidos dos dados brutos iniciais coletados de diversas fontes e passam por processo de limpeza e transformação necessárias para a próxima fase.

4. Modelagem: aqui nesta fase aplicam-se efetivamente as técnicas de modelagem visando à solução do problema proposto. Nesta fase são criados de modelos de machine learning tantos quantos forem necessários para uma avaliação de desempenho. Esta fase é iterativa inclusive pode ser necessário voltar à fase anterior para ajustar os dados.

5. Avaliação: realizam-se testes com o modelo escolhido que teve o melhor desempenho e valida-se para ver se atendem às necessidades do negócio. Nesta fase também há a possibilidade de se ajustar algum objetivo ou mesmo identificar se há algum novo objetivo que não tenha sido contemplado.

6. Utilização ou Aplicação: é nesta fase que a organização efetivamente aplica em seus processos diários toda a análise que foi gerada pelo modelo desenvolvido. Aqui nesta fase é onde são tomadas as decisões para a utilização ou não do modelo.

   Esta metodologia será aplicada durante o desnvolvimento deste trabalho.

2. Coleta de Dados

Para a realização deste trabalho, foram coletados 7 (sete) datasets disponibilizados por órgãos da administração pública federal, e portanto, são dados abertos e públicos. Segue o detalhamento dos mesmos.

2.1 – Dados Cadastrais das Pessoas Jurídicas.

Dataset obtido no link abaixo em janeiro de 2021.
https://receita.economia.gov.br/orientacao/tributaria/cadastros/cadastro-nacional-de-pessoas-juridicas-cnpj/dados-publicos-cnpj

Após o download dos vários arquivos, a base em arquivo formato csv foi obtida por meio do pacote R {qsacnpj} disponibilizado publicamente em :
https://github.com/georgevbsantiago/qsacnpj

DICIONÁRIO DE DADOS:

Nome da coluna/campo	Descrição	Tipo
CNPJ	CONTEM O NÚMERO DE INS-CRIÇÃO NO CNPJ (CADASTRO NACIONAL DA PESSOA JURÍDICA).	NUMÉRICO
IDENTIFICADOR MATRIZ/FILIAL	1 – MATRIZ 2 – FILIAL	NUMÉRICO

RAZÃO SOCIAL/NOME EMPRESARIAL	CORRESPONDE AO NOME EMPRESARIAL DA PESSOA JURÍDICA	CARACTER
NOME FANTASIA	CORRESPONDE AO NOME FANTASIA	CARACTER
SITUAÇÃO CADASTRAL	2 DIGITOS CÓDIGO DA SITUAÇÃO CADASTRAL 01 - NULA 02 - ATIVA 03 - SUSPENSA 04 - INAPTA 08 - BAIXADA	NUMÉRICO
DATA SITUACAO CADASTRAL	DATA DO EVENTO DA SITUACAO CADASTRAL	NUMÉRICO
MOTIVO SITUAÇÃO CADASTRAL	CÓDIGO DO MOTIVO DA SITUAÇÃO CADASTRAL	NUMÉRICO
NM-CIDADE EXTERIOR	NOME DA CIDADE NO EXTERIOR	CARACTER
CO-PAIS	CODIGO DO PAIS	CARACTER
NM-PAIS	NOME DO PAIS	CARACTER
CODIGO NATUREZA JURIDICA	CÓDIGO DA NATUREZA JURÍDICA	NUMÉRICO
DATA INICIO ATIVIDADE	DATA DE INICIO DA ATIVIDADE	NUMÉRICO
CNAE-FISCAL	INDICA O CÓDIGO DA ATIVIDADE ECONÔMICA PRINCIPAL DO ESTABELECIMENTO	NUMÉRICO
DESCRIÇÃO TIPO LOGRADOURO	CORRESPONDE A DESCRIÇÃO DO LOGRADOURO	CARACTER
LOGRADOURO	CORRESPONDE AO NOME DO LOGRADOURO ONDE SE LOCALIZA O ESTABELECIMENTO	CARACTER
NUMERO	CORRESPONDE AO NÚMERO ONDE SE LOCALIZA O ESTABELECIMENTO, QUANDO NÃO HOUVER PREENCHIMENTO DO NÚMERO HAVERÁ 'S/N'.	CARACTER
COMPLEMENTO	CORRESPONDE AO COMPLEMENTO PARA O ENDEREÇO DE LOCALIZAÇÃO DO ESTABELECIMENTO	CARACTER
BAIRRO	CORRESPONDE AO BAIRRO ONDE SE LOCALIZA O ESTABELECIMENTO	CARACTER
CEP	CÓDIGO DE ENDEREÇAMENTO POSTAL REFERENTE AO LOGRADOURO	NUMÉRICO

	NO QUAL O ESTABELECIMENTO ESTA LOCALIZADO	
UF	CORRESPONDE A SIGLA DA UNIDADE DA FEDERAÇÃO EM QUE SE ENCONTRA O ESTABELECIMENTO	CARACTER
CODIGO MUNICIPIO	CORRESPONDE AO CODIGO DO MUNICIPIO DE JURISDIÇÃO ONDE SE ENCONTRA O ESTABELECIMENTO	NUMÉRICO
MUNICIPIO	CORRESPONDE AO MUNICIPIO DE JURISDIÇÃO ONDE SE ENCONTRA O ESTABELECIMENTO	CARACTER
DDD-TELEFONE-1		CARACTER
DDD-1	DDD-1	CARACTER
TELEFONE-1	TELEFONE-1	CARACTER
DDD-TELEFONE-2		CARACTER
DDD-2	DDD-2	CARACTER
TELEFONE-2	TELEFONE-2	CARACTER
DDD-FAX		CARACTER
NU-DDD-FAX	DDD-FAX	CARACTER
NU-FAX	FAX	CARACTER
CORREIO ELETRONICO	E-MAIL DO CONTRIBUINTE	CARACTER
QUALIFICAÇÃO DO RESPONSÁVEL	QUALIFICAÇÃO DA PESSOA FÍSICA RESPONSÁVEL PELA EMPRESA	NUMÉRICO
CAPITAL SOCIAL DA EMPRESA	CAPITAL SOCIAL DA EMPRESA	NUMÉRICO
PORTE-EMPRESA	CÓDIGO DO PORTE DA EMPRESA 00 - NAO INFORMADO 01 - MICRO EMPRESA 03 - EMPRESA DE PEQUENO PORTE 05 - DEMAIS	CARACTER
OPÇÃO PELO SIMPLES	INDICADOR DA EXISTÊNCIA DA OPÇÃO PELO SIMPLES. 0 OU BRANCO - NÃO OPTANTE 5 E 7 – OPTANTESPELO SIMPLES 6 E 8 – EXCLUÍDO DO SIMPLES	CARACTER
DATA OPCAO PELO SIMPLES	DATA DE OPÇÃO PELO SIMPLES	NUMÉRICO
DATA EXCLUSÃO DO SIMPLES	DATA DE EXCLUSÃO DO SIMPLES	NUMÉRICO
OPÇÃO PELO MEI	INDICADOR DA EXISTÊNCIA DA OPÇÃO PELO MEI S – SIM N - NÃO	CARACTER

	OUTROS (BRANCO, ETC)	
SITUAÇÃO ESPECIAL	SITUAÇÃO ESPECIAL DA EMPRESA	CARACTER
DATA SITUAÇÃO ESPECIAL	DATA EM QUE A EMPRESA ENTROU EM SITUAÇÃO ESPECIAL (AAAAMMDD)	NUMÉRICO

2.2 – Dados dos sócios das Pessoas Jurídicas.

Dataset obtido no link abaixo em janeiro de 2021.

https://receita.economia.gov.br/orientacao/tributaria/cadastros/cadastro-nacional-de-pessoas-juridicas-cnpj/dados-publicos-cnpj

Após o download dos vários arquivos, a base em arquivo formato csv foi obtida por meio do pacote R {qsacnpj} disponibilizado publicamente em : https://github.com/georgevbsantiago/qsacnpj

DICIONÁRIO DE DADOS:

Nome da coluna/campo	Descrição	Tipo
CNPJ	CONTEM O NÚMERO DE INSCRIÇÃO NO CNPJ (CADASTRO NACIONAL DA PESSOA JURÍDICA).	NUMÉRICO
IDENTIFICADOR DE SOCIO	1 – PESSOA JURÍDICA 2 – PESSOA FISICA 3 – ESTRANGEIRO	NUMÉRICO
NOME SOCIO (NO CASO PF) OU RAZÃO SOCIAL (NO CASO PJ)	CORRESPONDE AO NOME SOCIO PESSOA FISICA, RAZÃO SOCIAL E/OU NOME EMPRESARIAL DA PESSOA JURÍDICA E NOME DO SÓCIO/RAZAO SOCIAL DO SOCIO ESTRANGEIRO	CARACTER
CNPJ/CPF DO SÓCIO	DADOS NÃO DISPONÍVEIS	NUMÉRICO
CODIGO QUALIFICACAO SOCIO	CODIGO QUALIFICACAO SOCIO	CARACTER
PERCENTUAL CAPITAL SOCIAL	ZEROS (VALORES NÃO CONSIDERADOS POR TER CARATER SIGILOSO)	NUMÉRICO
DATA ENTRADA SOCIEDADE	DATA DE ENTRADA NA SOCIEDADE	NUMÉRICO
CODIGO PAIS	CODIGO PAIS DO SOCIO ESTRANGEIRO(VALORES NÃO CONSIDERADOS)	CARACTER
NOME PAIS SOCIO	CORRESPONDE AO NOME DO PAIS DO	CARACTER

	SÓCIO(VALORES NÃO CONSIDERADOS)	
CPF REPRESENTANTE LEGAL	DADOS NÃO DISPONÍVEIS	NUMÉRICO
NOME REPRESENTANTE	DADOS NÃO DISPONÍVEIS	CARACTER
CODIGO QUALIFICACAO REPRESENTANTE LEGAL	CORRESPONDE AO CÓDIGO DA QUALIFICACAO DO REPRESENTANTE LEGAL	CARACTER

2.3 – Dados da qualificação dos sócios das Pessoas Jurídicas.

Dataset obtido no link abaixo em janeiro de 2021.

https://receita.economia.gov.br/orientacao/tributaria/cadastros/cadastro-nacional-de-pessoas-juridicas-cnpj/dados-publicos-cnpj

Esta tabela já estava disponível em formato de planilha.

DICIONÁRIO DE DADOS:

Nome da coluna/campo	Descrição	Tipo
CÓDIGO	CORRESPONDE AO CÓDIGO DA QUALIFICACAO DO REPRESENTANTE LEGAL	CARACTER
DESCRIÇÃO	CORRESPONDE A DESCRIÇÃO DA QUALIFICACAO DO REPRESENTANTE LEGAL	CARACTER
COLETADO ATUALMENTE	SIM / NÃO	CARACTER

2.4 – Dados da classificação da natureza jurídica das Pessoas Jurídicas.

Dataset obtido no link abaixo em janeiro de 2021.

https://concla.ibge.gov.br/estrutura/natjur-estrutura/natureza-juridica-2018

Outra alternativa é o link da Receita Federal:

https://receita.economia.gov.br/orientacao/tributaria/cadastros/cadastro-nacional-de-pessoas-juridicas-cnpj/tabelas-utilizadas-pelo-programa-cnpj/tabela-de-natureza-juridica-e-qualificacao-do-representante-da-entidade

DICIONÁRIO DE DADOS:

Nome da coluna/campo	Descrição	Tipo
CÓDIGO NATUREZA JURÍDICA	CORRESPONDE AO CÓDIGO DA NATUREZA JURIDICA	NUMERICO
NOME NATUREZA JURÍDICA	CORRESPONDE A DESCRIÇÃO DA NATUREZA JURIDICA	CARACTER
CODIGO SUCLASS NATUREZA JURÍDICA	CODIGO DA SUCLASSIFICAÇÃO DA NATUREZA JURÍDICA	NUMERICO
NOME DA SUBCLASS NATURE-ZA JURÍDICA	DESCRIÇÃO DA SUBCLASSIFI-CAÇÃO DA NATUREZA JURÍDICA	CARACTER

2.5 – Dados do CNAE 2.3 das Pessoas Jurídicas.

Dataset obtido no link abaixo em janeiro de 2021.

https://concla.ibge.gov.br/classificacoes/por-tema/atividades-economicas

DICIONÁRIO DE DADOS:

Nome da coluna/campo	Descrição	Tipo
CÓDIGO SEÇÃO	CORRESPONDE AO CÓDIGO DA SEÇÃO	CARACTER
NOME SEÇÃO	DESCRIÇÃO DO CODIGO DA SEÇÃO	CARACTER
CODIGO DIVISÃO	CÓDIGO DA DIVISÃO	NUMERICO
NOME DIVISAO	DESCRIÇÃO DO CODIGO DA DIVISÃO	CARACTER
CODIGO GRUPO	CODIGO DO GRUPO	CARACTER
NOME GRUPO	DESCRIÇÃO DO CÓDIGO DO GRUPO	CARACTER
CODIGO CLASSE	CODIGO DA CLASSE	CARACTER
NOME CLASSE	DESCRIÇÃO DO CODIGO DA CLASSE	CARACTER
CODIGO CNAE	CÓDIGO DA CNAE	CARACTER
NOME CNAE	DESCRIÇÃO DO CÓDIGO DO CNAE	CARACTER

2.6 – Dados de compras efetuadas pela Administração Pública Federal

Dataset obtido no link abaixo em janeiro de 2021.

http://portaltransparencia.gov.br/download-de-dados/compras

Foram baixadas as planilhas relativas ao período de janeiro de 2020 a novembro de 2020.

DICIONÁRIO DE DADOS:

Nome da coluna/campo	Descrição	Tipo
NÚMERO DO CONTRATO	NÚMERO QUE IDENTIFICA O CONTRATO NO SIASG	CARACTER
OBJETO	OBJETO DO CONTRATO	CARACTER
FUNDAMENTO LEGAL	INDICAÇÃO DO EMBASAMENTO LEGAL DO CONTRATO	CARACTER
MODALIDADE DE COMPRA	CONCORRÊNCIA CONCURSO; CONVITE; DISPENSA DE LICITAÇÃO; INEXIGIBILIDADE DE LICITAÇÃO; PREGÃO; REGISTRO DE PREÇO; TOMADA DE PREÇOS.	CARACTER
SITUAÇÃO CONTRATO	SITUAÇÃO EM QUE SE ENCONTRA O CONTRATO	CARACTER
CÓDIGO DO ÓRGÃO SUPERIOR	CÓDIGO DO ÓRGÃO SUPERIOR RESPONSÁVEL PELA LICITAÇÃO ÓRGÃO SUPERIOR - UNIDADE DA ADMINISTRAÇÃO DIRETA QUE TENHA ENTIDADES POR ELE SUPERVISIONADAS. *FONTE: MANUAL DO SIAFI*	CARACTER
NOME ÓRGÃO SUPERIOR	NOME DO ÓRGÃO SUPERIOR	CARACTER
CÓDIGO ÓRGÃO	CÓDIGO DO ÓRGÃO RESPONSÁVEL PELA LICITAÇÃO ÓRGÃO SUBORDINADO - ENTIDADE SUPERVISIONADA POR UM ÓRGÃO DA ADMINISTRAÇÃO DIRETA. *FONTE: MANUAL DO SIAFI*	CARACTER
NOME ÓRGÃO	NOME DO ÓRGÃO	CARACTER
CÓDIGO UG	CÓDIGO DA UNIDADE GESTORA RESPONSÁVEL PELA LICITAÇÃO. UNIDADE GESTORA (UG) - UNIDADE ORÇAMENTÁRIA OU ADMINISTRATIVA QUE REALIZA ATOS DE GESTÃO ORÇAMENTÁRIA, FINANCEIRA E/OU PATRIMONIAL, CUJO TITULAR, EM CONSEQUÊNCIA, ESTÁ SUJEITO A TOMADA DE CONTAS ANUAL NA CONFORMIDADE DO DISPOSTO NOS ARTIGOS 81 E 82 DO DECRETO-LEI NR. 200, DE 25 DE FEVEREIRO DE 1967. *FONTE: MANUAL DO SIAFI*	CARACTER
NOME UG	NOME DA UNIDADE GESTORA	CARACTER
DATA ASSINATURA CONTRATO	DATA DA ASSINATURA DO CONTRATO	CARACTER
DATA PUBLICAÇÃO DOU	DATA DA PUBLICAÇÃO DO CONTRATO NO DOU	CARACTER
DATA INÍCIO DA VIGÊNCIA	DATA DE INÍCIO DA VIGÊNCIA DO CONTRATO	CARACTER
DATA FIM DA VIGÊNCIA	DATA DE FIM DA VIGÊNCIA DO CONTRATO	CARACTER
CNPJ CONTRATADO	CNPJ DO CONTRATADO	CARACTER

NOME CONTRATADO	NOME DO CONTRATADO	CARACTER
VALOR INICIAL DA COMPRA	VALOR INICIAL DA COMPRA	NUMÉRICO
VALOR FINAL DA COMPRA	VALOR FINAL DA COMPRA	NUMÉRICO

2.7 – Dados das Pessoas Jurídicas impedidas de contratar com a Administração Pública Federal (Empresas Inidôneas)

Dataset obtido no link abaixo em janeiro de 2021.

http://www.portaltransparencia.gov.br/download-de-dados/ceis

Os dados disponíveis estavam atualizados até 16 de janeiro de 2021.

DICIONÁRIO DE DADOS:

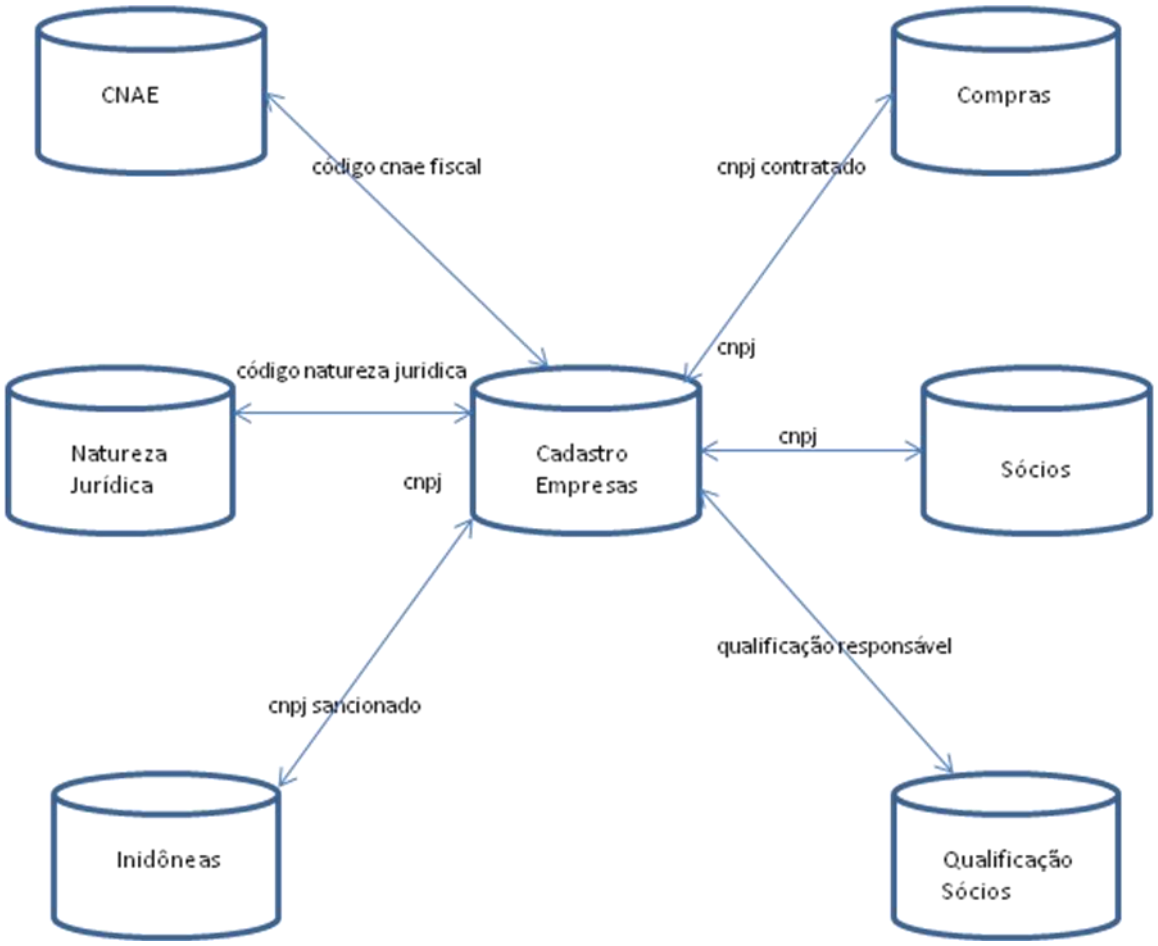
Nome da coluna/campo	Descrição	Tipo
TIPO DE PESSOA	IDENTIFICA SE A PENALIDADE FOI APLICADA A "PESSOA FÍSICA" OU "PESSOA JURÍDICA".	CARACTER
CPF OU CNPJ DO SANCIONADO	NÚMERO DE CADASTRO DO SANCIONADO JUNTO À RECEITA FEDERAL: CPF PARA PESSOAS FÍSICAS E CNPJ PARA PESSOAS JURÍDICAS.	CARACTER
NOME INFORMADO PELO ÓRGÃO SANCIONADOR	CONFORME REGISTRADO PELO ÓRGÃO SANCIONADOR NO SIRCAD, OU CONFORME PUBLICADO NO DOU.	CARACTER
RAZÃO SOCIAL – CADASTRO RECEITA	CAMPO EXTRAÍDO DA BASE CNPJ OU DA BASE CPF DA RECEITA FEDERAL (RESULTADO DA BUSCA PELO VALOR DO CAMPO "CPF OU CNPJ DO SANCIONADO")	CARACTER
NOME FANTASIA – CADASTRO RECEITA	IDEM ANTERIOR. VALE REGISTRAR QUE AS INFORMAÇÕES DE IDENTIFICAÇÃO DO SANCIONADO (NOME INFORMADO, RAZÃO SOCIAL E NOME FANTASIA) SÃO MANTIDAS NO CEIS PARA FACILITAR A PESQUISA E DAR TRANSPARÊNCIA ÀS SANÇÕES QUANDO OCORRE MUDANÇA DE ALGUM DESTES DADOS DO SANCIONADO.	CARACTER
NÚMERO DO PROCESSO	NÚMERO DO PROCESSO NO ÂMBITO DO QUAL FOI APLICADA A SANÇÃO.	CARACTER
TIPO SANÇÃO	TODAS AS SANÇÕES QUE IMPLIQUEM EM RESTRIÇÃO AO DIREITO DE PARTICIPAR DE LICITAÇÕES OU DE CELEBRAR CONTRATOS COM A ADMINISTRAÇÃO PÚBLICA. O ROL DE SANÇÕES E RESPECTIVAS FUNDAMENTAÇÕES LEGAIS ESTÁ DISPONÍVEL EM HTTP://WWW.PORTALDATRANSPARENCIA.GO	CARACTER

	V.BR/CEIS/SAIBA-MAIS	
DATA INÍCIO SANÇÃO	CONSIDERA-SE A DATA DA PUBLICAÇÃO DA SANÇÃO, QUANDO NÃO HOUVER MENÇÃO EXPRESSA À DATA DE INÍCIO DE VIGÊNCIA DA PENALIDADE.	CARACTER
DATA FINAL SANÇÃO	CONSIDERA-SE O PRAZO ESTABELECIDO PARA O TÉRMINO DE VIGÊNCIA DA PENALIDADE. NO CASO DA DECLARAÇÃO DE INIDONEIDADE, MESMO QUE CONSTE UM PRAZO DE VIGÊNCIA ESTE É CONSIDERADO PRAZO MÍNIMO DA PENALIDADE. PORTANTO A INIDONEIDADE SÓ É EXCLUÍDA DO CEIS MEDIANTE INFORMAÇÃO DA REABILITAÇÃO DO SANCIONADO (PUBLICAÇÃO DA REABILITAÇÃO NO DOU, REGISTRO NO SIRCAD E DEMAIS BASES OU APRESENTAÇÃO DA DECISÃO PELA REABILITAÇÃO).	CARACTER
ÓRGÃO SANCIONADOR	ÓRGÃO ESPECÍFICO QUE APLICOU A SANÇÃO.	CARACTER
UF ÓRGÃO SANCIONADOR	UNIDADE DA FEDERAÇÃO DO ÓRGÃO RESPONSÁVEL PELA APLICAÇÃO DA SANÇÃO.	CARACTER
ORIGEM INFORMAÇÕES	ÓRGÃO QUE INFORMOU A SANÇÃO. HÁ, POR EXEMPLO, GOVERNOS ESTADUAIS QUE ESTABELECEM UM ÓRGÃO COMO O RESPONSÁVEL PELO REGISTRO DAS SANÇÕES APLICADAS POR TODOS OS ENTES DAQUELE GOVERNO. TAMBÉM É O CASO DO CNJ, QUE MANTÉM CADASTRO DAS SANÇÕES APLICADAS POR TODOS OS ÓRGÃOS JUDICIÁRIOS.	CARACTER
DATA ORIGEM INFORMAÇÕES	DATA DE REGISTRO DA SANÇÃO NO CEIS.	CARACTER
DATA PUBLICAÇÃO	DATA DA PUBLICAÇÃO DA SANÇÃO EM VEÍCULO OFICIAL DE INFORMAÇÃO.	CARACTER
PUBLICAÇÃO	VEÍCULO OFICIAL DE INFORMAÇÃO ONDE A SANÇÃO FOI PUBLICADA.	CARACTER
DETALHAMENTO	DADOS DA PUBLICAÇÃO, COMO POR EXEMPLO A SEÇÃO E A PÁGINA DO DOU.	CARACTER
ABRAGÊNCIA DEFINIDA EM DECISÃO JUDICIAL	O CAMPO SÓ É PREENCHIDO QUANDO HÁ DETERMINAÇÃO PELA JUSTIÇA DA ABRANGÊNCIA DA SANÇÃO. NOS DEMAIS CASOS, A INTERPRETAÇÃO QUANTO À ABRANGÊNCIA DA SANÇÃO É DE RESPONSABILIDADE DO USUÁRIO DO CADASTRO.	CARACTER
FUNDAMENTAÇÃO LEGAL	DISPOSITIVO LEGAL QUE FUNDAMENTA A APLICAÇÃO DA SANÇÃO.	CARACTER

DESCRIÇÃO DA FUNDAMENTAÇÃO LEGAL	DETALHAMENTO DA NORMA QUE FUNDAMENTA A APLICAÇÃO DA SANÇÃO.	CARACTER
DATA DO TRÂNSITO EM JULGADO	CAMPO OPCIONAL QUE INDICA A DATA EM QUE A DECISÃO JUDICIAL PELA APLICAÇÃO DA SANÇÃO TRANSITOU EM JULGADO, OU SEJA, QUANDO NÃO SE PODE MAIS RECORRER JUDICIALMENTE DESTA DECISÃO.	CARACTER
COMPLEMENTO DO ÓRGÃO	CAMPO OPCIONAL QUE DETALHA, QUANDO PERTINENTE, A UNIDADE RESPONSÁVEL PELA APLICAÇÃO DA SANÇÃO. TRATA-SE DE DETALHAMENTO DA INFORMAÇÃO DA COLUNA "ÓRGÃO SANCIONADOR"	CARACTER

Na figura abaixo observa-se um diagrama com as relações entre as tabelas de dados que serão utilizadas para a criação da tabela de análises e criação do modelo.

Não é um diagrama lógico de modelo de dados de E-R, é apenas para se ter uma visão macro dos relacionamentos entre as tabelas.



3. Processamento/Tratamento de Dados

Uma vez coletados e reunidos os dados em arquivos csv, o próximo passo foi utilizar uma ferramenta que permitisse trabalhar estas tabelas em um único local para que fosse possível executar pesquisas, selecionar dados, transformar, agrupar, limpar, criar novos dados e ao final montar uma tabela que estivesse pronta para ser utilizada no processo de análise e criação de um modelo estatístico.

Alguns destes arquivos possuem tamanho considerável, conforme pode-se verificar na tabela abaixo abaixo:

Nome arquivo csv	Descrição	Tamanho
CNPJ_DADOS_CADASTRAIS_PJ.CSV	CADASTRO DAS PESSOAS JURÍDICAS	10.3 GB
CNPJ_DADOS_SOCIOS_PJ.CSV	DADOS DOS SÓCIOS DAS PESSOAS JURÍDICAS	3.0 GB
INIDÔNEAS.CSV	INFORMAÇÕES DAS EMPRESAS IMPEDIDAS	25.7 MB
COMPRAS.CSV	COMPRAS EFETUADAS	9.5 MB
TAB_CNAE.CSV	CÓDIGOS E DESCRIÇÃO CNAE	372 KB
TAB_NATUREZA_JURIDICA.CSV	CÓDIGOS E DESCRIÇÃO NATUREZA JURÍDICA	8 KB
TAB_QUALIFICACAO_RESPONSAVEL_SOCIO.CSV	CODIGOS E DESCRIÇÃO DA QUALIFICAÇÃO DOS RESPONSÁVEIS	4 KB

Para estas atividades, foi escolhida a ferramenta de banco de dados MySQL Community Server, de licença GPL (General Public License) que pode ser obtido no link: https://dev.mysql.com/downloads/mysql/ juntamente com a ferramenta MySQL Workbench, também de licença GPL, que pode ser obtida no link: https://dev.mysql.com/downloads/workbench/

A escolha desta ferramenta levou em consideração os seguintes pontos:

1. Apropriada para se trabalhar com tabelas grandes e dados estruturados;
2. Flexibilidade e facilidade de uso para cruzamento de dados entre as tabelas;

3. Uso da linguagem SQL, que eu domino por já trabalhar a algum tempo realizando atividades de manipulação de dados e é a linguagem que eu utilizo no dia-a-dia para execução dos meus projetos de Data Science.

Todos os scripts em SQL que foram utilizados neste trabalho serão anexados a este documento na seção destinada para este fim.

A primeira etapa foi realizar a importação dos dados em tabelas seguindo as informações documentadas nos dicionários de dados com respeito aos campos de dados e seus tipos.

Em seguida, foi inciado o trabalho de junção das tabelas com objetivo de construir uma ABT (Analytical Base Table). A ABT é uma tabela de dados que é criada com base nas consultas efetuadas a diversas fontes de dados. Normalmente essas consultas são obtidas em bancos de dados (estruturados ou não estruturados), em Data Lakes, em DataWarehouses ou mesmo por Web Scraping. Ao final, esta tabela ABT é utilizada para a realização de análises estatísticas bem como para alimentar modelos de machine learning, gerar informações e chegar a conclusões para os problemas propostos.

Na grande maioria das vezes as respostas aos problemas não estão em uma única tabela em um banco de dados ou em um único local na web ou em uma única fonte de dados. Por esta razão, com a utilização de ferramentas como a linguagem SQL, consegue-se obter as informações desejadas e ao mesmo tempo já realiza atividades de transformação dos dados. A propósito, nesta etapa é muito importante a participação dos responsáveis pelo negócio, pois eles podem dar orientações de quais informações são importantes e devem ser usadas para alimentar o modelo e assim apresentar soluções aos problemas.

Ao fim desta etapa de processamento dos dados, foi criada a seguinte ABT:

DICIONÁRIO DE DADOS DA ABT:

Nome da coluna/campo	Descrição	Tipo
CNPJ	CNPJ DA PESSOA JURIDICA	CARACTER
IDENTIFICADOR_MATRIZ_FILIAL	IDENTIFICA SE É MATRIZ OU FILIAL	CARACTER
RAZAO_SOCIAL	RAZÃO SOCIAL DA PESSOA JURÍDICA	CARACTER

SITUAÇÃO_CADASTRAL	CODIGO DA SITUAÇÃO CADASTRAL	CARACTER
DATA_SITUACAO_CADASTRAL	DATA DA SITUAÇÃO CADASTRAL	DATE
MOTIVO_SITUACAO_CADASTRAL	CODIGO DO MOTIVO DA SITUAÇÃO CADASTRAL	CARACTER
CODIGO_NATUREZA_JURIDICA	CÓDIGO DA NATUREZA JURÍDICA	CARACTER
NM_SUBCLAS_NAT_JUR	DESCRIÇÃO DA NATUREZA JURÍDICA	CARACTER
DATA_INICIO_ATIVIDADE	DATA DE INICIO DAS ATIVIDADES DA PESSOA JURÍDICA	DATE
IDADE_EMPRESA_MESES	VARIÁVEL CALCULADA. INDICA A IDADE DA EMPRESA EM MESES	NUMERICO
CNAE_FISCAL	CODIGO DO CNAE FISCAL	CARACTER
NM_CNAE	DESCRIÇÃO DO CNAE FISCAL	CARACTER
CEP	CEP DA LOCALIZAÇÃO DA PESSOA JURÍDICA	CARACTER
UF	UNIDADE DA FEDERAÇÃO	CARACTER
CODIGO_MUNICIPIO	CÓDIGO DO MUNICÍPIO	CARACTER
MUNICIPIO	DESCRIÇÃO DO MUNICÍPIO	CARACTER
QUALIFICACAO_DO_RESPONSAVEL	CODIGO DA QUALIFICAÇÃO DO RESPONSÁVEL	NUMERICO
NOME_QUALIF_SOCIO	DESCRIÇÃO DA QUALIFICAÇÃO DO RESPONSÁVEL	CARACTER
CAPITAL_SOCIAL	CAPITAL SOCIAL	NUMERICO
PORTE	CODIGO DO PORTE DA EMPRESA	CARACTER
NM_PORTE	DESCRIÇÃO DO PORTE DA EMPRESA	CARACTER
OPCAO_PELO_SIMPLES	CODIGO DA OPÇÃO PELO SIMPLES	CARACTER
NM_OPTANTE_SIMPLES	DESCRIÇÃO DA OPÇÃO PELO SIMPLES	CARACTER
OPCAO_PELO_MEI	CODIGO DA OPÇÃO PELO MEI	CARACTER
ULTIMA_DATA_ENTRADA	VARIÁVEL CALCULADA ULTIMA DATA DE ENTRADA DE ALGUM SOCIO NA PESSOA JURÍDICA	DATE
QTDE_MESES_ENTR_SOCIO	VARIÁVEL CALCULADA QUANTIDADE D MESES DESDE A ULTIMA ENTRADA DE ALGUM SOCIO	NUMERICO
NRO_SOCIOS	VARIÁVEL CALCULADA QUANTIDADE DE SÓCIOS	NUMERICO
MODALIDADE_COMPRA	DESCRIÇÃO DA MODALIDADE DE COMPRA	CARACTER
OBJETO	DESCRIÇÃO DO OBJETO DA CONTRATAÇÃO	
CODIGO_ORGAO	CÓDIGO DO ORGÃO CONTRATANTE	CARACTER
NOME_ORGAO	DESCRIÇÃO DO ÓRGÃO CONTRATANTE	CARACTER
CODIGO_UG	CÓDIGO DA UNIDADE GESTORA DO CONTRATO	CARACTER
NOME_UG	NOME DA UNIDADE GESTORA DO CONTRATO	CARACTER
VALOR_INICIAL_COMPRA	VALOR INICIAL DO CONTRATO	NUMERICO
VALOR_FINAL_COMPRA	VALOR FINAL DO CONTRATO	NUMERICO
DIFERENCA_COMPRA	VARIÁVEL CALCULADA DIFERENÇA ENTRE O VALOR FINAL	NUMERICO

	E O VALOR INCIAL	
AUMENTO_VALOR_CONTRATO	VARIAVEL CRIADA INDICA SE HOUVE AUMENTO DE VALOR NO CONTRATO	CARACTER
INIDONEA	VARIÁVEL TARGET INDICA DE A PESSOA JURÍDICA ESTÁ IMPEDIDA DE CONTRATAR	NUMERICO

A tabela possui 18.254 observações e 38 variáveis.

Algumas análises foram feitas em alguns campos para verificar a possível utilidade da variável no desenvolvimento de um modelo. Além disso, foi utilizada a técnica de engenharia de atributos para a criação de novas variáveis que possam ser úteis para a modelagem.

1. Variável: situacao_cadastral

   Esta variável possui os seguintes valores e quantitativos:

situacao_cadastral	count(*)
02	18247
08	5
03	2

   Será removida da análise pois não tem representatividade nos dados pois tem alta duplicidade.

2. Variável: motivo_situacao_cadastral

   Pelo mesmo motivo, esta variável também não será utilizada:

motivo_situacao_cadastral	count(*)
00	18247
02	3
01	2
36	2

   Também removida da análise pois não tem representatividade nos dados pois tem alta duplicidade.

3. Variável: data_situacao_cadastral

   Esta variável está relacionada com as informações das variáveis *situacao_cadastral* e *motivo_situacao_cadastral* e portanto não há motivo para manter na análise.

4. Variável: objeto

   Esta variável é uma descrição detalhada do objeto da contratação:

objeto
Objeto: Aquisição de material de uso laboratorial.
Objeto: Aquisição de material de uso laboratorial.
Objeto: Serviço de construção civil para execução da adequação de espaços do Bloco I do Campus Presidente Epitácio.
Objeto: Aquisição de equipamentos de proteção individual - EPI para motociclistas (Joelheira, Luva e Baladava), conforme condições, quantidades e exigências estabelecidas no Termo de Referência e Anexos ...
Objeto: Contratação de empresa especializada na prestação de serviços de informática na modalidade de FÁBRICA DE SOFTWARE.
Objeto: Contratação de empresa especializada na prestação de serviços de informática na modalidade de FÁBRICA DE SOFTWARE.
Objeto: Aquisição de materiais para uso laboratorial.
Objeto: Aquisição de material de uso laboratorial.
Objeto: Aquisição de materiais para uso laboratorial.
Objeto: Aquisição de materiais para uso laboratorial.
Objeto: Aquisição de kits, reagentes, ensaios, anticorpos, etc.
Objeto: Aquisição de ônibus rodoviário.
Objeto: Aquisição de veículos para utilização nas atividades das Organizações Militares da Força Aérea Brasileira.
Objeto: Aquisição de material de uso laboratorial.
Objeto: Aquisição de veículos para utilização nas atividades das Organizações Militares da Força Aérea Brasileira.
Objeto: Contratação de empresa especializada para prestação de serviços de impressão (outsoursing de impressão) na modalidade franquia de páginas mais excedente pelo prazo de 48 meses.
Objeto: Contratação de empresa especializada para prestação de serviços de impressão (outsoursing de impressão) na modalidade franquia de páginas mais excedente pelo prazo de 48 meses.
Objeto: Aquisição de equipamentos de proteção individual - EPI para motociclistas (Calça e Jaqueta)
Objeto: Aquisição de material de uso laboratorial.
Objeto: Aquisição de veículos para utilização nas atividades das Organizações Militares da Força Aérea Brasileira.
Objeto: Aquisição de veículo zero quilômetro: Cavalo Mecânico especial para reboque trailer com a finalidade de funcionamento de Unidade Médico - Odontológica.
Objeto: Prestação de serviços de tecnologia da informação para manutenção evolutiva, perfectiva e adaptativa de sistemas de informação e de serviços técnicos de tecnologia da informação para sustentação ...
Objeto: Prestação de serviços de tecnologia da informação para manutenção evolutiva, perfectiva e adaptativa de sistemas de informação e de serviços técnicos de tecnologia da informação para sustentação ...
Objeto: Prestação de serviços de tecnologia da informação para manutenção evolutiva, perfectiva e adaptativa de sistemas de informação e de serviços técnicos de tecnologia da informação para sustentação ...
Objeto: Prestação de serviços de tecnologia da informação para manutenção evolutiva, perfectiva e adaptativa de sistemas de informação e de serviços técnicos de tecnologia da informação para sustentação ...
Objeto: Aquisição de focos cirúrgicos.
Objeto: Contratação de serviços de pintura para readequação das instalações da Companhia de Manutenção para recebimento do Programa Estratégico do Exército Guarani
OBJETO: Contratação de obra de engenharia da reforma e ampliação da Capela
Objeto: Contratação de serviços relativos ao Grupo de Qualidade de Sistemas e Arquitetura de Softwares.
Objeto: Contratação de serviços relativos ao Grupo de Qualidade de Sistemas e Arquitetura de Softwares.
Objeto: Contratação de serviços relativos ao Grupo de Qualidade de Sistemas e Arquitetura de Softwares.
Objeto: Contratação de serviços relativos ao Grupo de Qualidade de Sistemas e Arquitetura de Softwares.
Objeto: Aquisição de Finasterida 5mg, Fluoxetina 20mg e Glicerol 12%

Por ser um campo de texto livre, torna-se bastante apropriada para uma possível modelagem de Text Mining. Desta forma como este não é o objetivo do presente trabalho será também excluído da análise.

5. Variável: idade_empresa_meses

   Esta variável foi criada a partir de outra variável no dataset utilizando a seguite fórmula:

   idade_empresa_meses = data_atual - data_inicio_atividade

   Ou seja, realizando cálculo de datas no SQL calculando a diferença em meses entre a data atual (do sistema) e a data do início das atividades da empresa.

   Esta variável foi obtida a partir da experiência do negócio.

6. Variável: ultima_data_entrada

Esta variável foi criada utilizando as informações dos sócios das empresas e as datas em que eles entraram na sociedade. Foi utilizada a função *Max* do SQL para calcular a data mais recente de entrada de um sócio.

ultima_data_entrada = max(data_entrada_sociedade)

A variável *data_entrada_sociedade* é uma coluna da tabela *sócios*.

7. Variável: qtde_meses_entr_socio

   A partir da variável anterior, foi calculada a diferença entre a data atual e a data de entrada do sócio:

   qtde_meses_entr_socio = max(data_entrada_sociedade) – data_atual

   Esta variável foi obtida a partir da experiência do negócio.

8. Variável: nro_socios

   Esta variável foi criada para representar a quantidade de sócios existentes nas pessoas jurídicas. Foi calculada a partir da função de contagem do SQL combinada com o agrupamento por pessoa jurídica, na tabela *sócios*:

   count(*) as nro_socios

   Esta variável foi obtida a partir da experiência do negócio.

9. Variável: diferenca_compra

   Esta variável foi criada calculando-se a diferença entre o valor final do contrato e o valor inicial do contrato:

   diferenca_compra = valor_final_compra - valor_inicial_compra

10. Variável: aumento_valor_contrato

    Esta variável categórica é um indicador se houve um aumento no valor do contrato e foi utilizada a variável *diferenca_compra* para obtê-la. Representa ‘S’ caso o valor da variável *diferenca_compra* seja maior do que zero e ‘N’ caso seja menor ou igual a zero.

Todos os cálculos utilizados na criação das variáveis acima descritas estão documentados nos scripts SQL que serão anexados a este trabalho.

4. Análise e Exploração dos Dados

A Análise Exploratória dos Dados ou simplesmente AED é uma abordagem para análise de conjuntos de dados de modo a resumir as suas características principais frequentemente não só com métodos visuais mas também utilizando-se métodos não visuais. Um dos principais objetivos da AED é observar o que os dados podem nos dizer antes de se aplicar qualquer algoritmo de modelagem. A AED emprega grande variedade de técnicas gráficas e quantitativas visando facilitar a obtenção de informações ocultas nos dados, descobrir variáveis importantes, detectar comportamentos anômalos, escolher modelos a serem usados e determinar o número ótimo de variáveis.

O dataset a ser analisado possui 18.234 observações e 34 variáveis sendo uma delas a variável resposta. A seguir uma visualização geral das estatísticas destas variáveis.

```
---------------------------------------------------------------------------
identificador_matriz_filial
       n  missing distinct
   18254        0        2
Value          1     2
Frequency  16978  1276
Proportion  0.93  0.07
---------------------------------------------------------------------------
codigo_natureza_juridica
       n  missing distinct
   18254        0       29
lowest : 1015 1023 1031 1104 1112, highest: 3077 3085 3131 3220 3999
---------------------------------------------------------------------------
nm_subclas_nat_jur
       n  missing distinct
   18254        0       29
lowest : Associação Privada                    Autarquia Estadual ou do Distrito Federal Autarquia Federal
Autarquia Municipal                Clube/Fundo de Investimento
```

```
highest: Sociedade de Economia Mista          Sociedade Empresária Limitada          Sociedade Simples
Limitada          Sociedade Simples Pura          Sociedade Unipessoal de Advogados
---------------------------------------------------------------------------
idade_empresa_meses
       n  missing distinct     Info     Mean      Gmd      .05      .10
   18254        0      647        1    224.4    175.5       36       51
     .25      .50      .75      .90      .95
     103      183      293      474      611
lowest :    6    7    8    9   10, highest:  875  935  985  992 1357
---------------------------------------------------------------------------
cnae_fiscal
       n  missing distinct     Info     Mean      Gmd      .05      .10
   18254        0      689        1  5710495  2302532  2621300  3250705
     .25      .50      .75      .90      .95
 4399103  4763602  7911200  8610102  8640210
lowest :  113000  116499  119999  121101  155505
highest: 9603304 9603399 9609207 9609299 9700500
---------------------------------------------------------------------------
cep
       n  missing distinct     Info     Mean      Gmd      .05      .10
   18254        0     7602        1 49805783 32894250  4575684  7040026
     .25      .50      .75      .90      .95
22775003 58015445 72110150 86807587 90423571
lowest :  1008000  1009000  1010000  1014000  1037000
highest: 99740000 99830000 99870000 99900000 99930000
---------------------------------------------------------------------------
uf
       n  missing distinct
   18254        0       27
lowest : AC AL AM AP BA, highest: RS SC SE SP TO
---------------------------------------------------------------------------
codigo_municipio
       n  missing distinct     Info     Mean      Gmd      .05      .10
   18254        0     1019    0.998     5754     3329      301      990
     .25      .50      .75      .90      .95
    3849     6001     8105     9701     9701
lowest :    1    3    4    5    7, highest: 9891 9907 9923 9951 9983
---------------------------------------------------------------------------
municipio
       n  missing distinct
   18254        0     1010
lowest : ABADIA DE GOIAS ABAETETUBA       ABREU E LIMA     ACAIACA          ACOPIARA
highest: VOLTA REDONDA   VOTORANTIM       VOTUPORANGA      XANXERE          XIQUE-XIQUE
---------------------------------------------------------------------------
qualificacao_do_responsavel
```

```
      n  missing distinct
  18254        0       10
lowest : 5  10 12 16 19, highest: 43 49 50 64 65
Value          5    10    12    16    19    43    49    50    64    65
Frequency   1563  1018     2  1896     1     1  8573  1234   179  3787
Proportion 0.086 0.056 0.000 0.104 0.000 0.000 0.470 0.068 0.010 0.207
---------------------------------------------------------------------------
capital_social
       n   missing  distinct      Info      Mean       Gmd       .05
   18254         0      1441     0.996 774357855 1.517e+09  0.00e+00
     .10       .25       .50       .75       .90       .95
 0.00e+00  3.30e+04  2.00e+05  1.28e+06  2.11e+07  4.27e+08
lowest :           0           1         100         110         250
highest: 17567609121 18744414943 32038471375 63571415865 79100000000
0 (17372, 0.952), 1e+09 (274, 0.015), 2e+09 (73, 0.004), 3e+09 (38,
0.002), 4e+09 (3, 0.000), 5e+09 (43, 0.002), 6e+09 (20, 0.001), 7e+09
(19, 0.001), 8e+09 (1, 0.000), 1.1e+10 (7, 0.000), 1.3e+10 (25, 0.001),
1.5e+10 (5, 0.000), 1.6e+10 (25, 0.001), 1.8e+10 (68, 0.004), 1.9e+10
(98, 0.005), 3.2e+10 (83, 0.005), 6.4e+10 (99, 0.005), 7.9e+10 (1,
0.000)
For the frequency table, variable is rounded to the nearest 1e+09
---------------------------------------------------------------------------
porte
      n  missing distinct     Info     Mean      Gmd
  18254        0        3     0.86    3.421    1.758
Value          1     3     5
Frequency   4673  5062  8519
Proportion 0.256 0.277 0.467
---------------------------------------------------------------------------
nm_porte
      n  missing distinct
  18254        0        3
Value                  Demais Empresa Pequeno Porte
Frequency                   8519                5062
Proportion                 0.467               0.277
Value             Micro Empresa
Frequency                  4673
Proportion                 0.256
---------------------------------------------------------------------------
opcao_pelo_simples
      n  missing distinct
  18254        0        5
lowest : 0 5 6 7 8, highest: 0 5 6 7 8
Value          0     5     6     7     8
Frequency   8457  5855  2745   778   419
```

Proportion 0.463 0.321 0.150 0.043 0.023

---

nm_optante_simples

n	missing	distinct
18254	0	3

Value	Excluido	Não Optante	Optante
Frequency	3164	8457	6633
Proportion	0.173	0.463	0.363

---

opcao_pelo_mei

n	missing	distinct
16978	1276	2

Value	N	S
Frequency	16866	112
Proportion	0.993	0.007

---

qtde_meses_entr_socio

n	missing	distinct	Info	Mean	Gmd	.05	.10
16924	1330	412	1	81.68	82.2	8	12

.25	.50	.75	.90	.95
22	52	113	197	253

lowest : 5 6 7 8 9, highest: 501 511 518 545 609

---

nro_socios

n	missing	distinct	Info	Mean	Gmd	.05	.10
16924	1330	47	0.919	2.962	2.494	1	1

.25	.50	.75	.90	.95
1	2	3	6	7

lowest : 1 2 3 4 5, highest: 166 171 181 254 362

---

modalidade_compra

n	missing	distinct
18254	0	9

lowest : Concorrência Concorrência - Registro de Preço Convite Dispensa de Licitação Inexigibilidade de Licitação

highest: Inexigibilidade de Licitação Pregão Pregão - Registro de Preço Sem Informação Tomada de Preços

Concorrência (78, 0.004), Concorrência - Registro de Preço (9, 0.000), Convite (34, 0.002), Dispensa de Licitação (3672, 0.201), Inexigibilidade de Licitação (2438, 0.134), Pregão (4121, 0.226), Pregão - Registro de Preço (5902, 0.323), Sem Informação (1685, 0.092), Tomada de Preços (315, 0.017)

---

codigo_orgao

n	missing	distinct	Info	Mean	Gmd	.05	.10

```
  18254       0     218    0.993    36318    13406    22203    25000
    .25      .50      .75      .90      .95
  26282    30911    52111    52121    52131
lowest : 15000 20101 20116 20202 20203, highest: 68201 81000 91081 95320 97400
---------------------------------------------------------------------------
nome_orgao
      n  missing distinct
  18254       0     219
lowest : Advocacia-Geral da União - Unidades com víncu Agência Nacional de Águas                Agência Nacional de Aviação Civil           Agência Nacional de Energia Elétrica        Agência Nacional de Mineração
highest: Universidade Federal Rural de Pernambuco       Universidade Federal Rural do Rio de Janeiro Universidade Federal Rural do Semi-Árido        Universidade Tecnológica Federal do Paraná  VALEC Engenharia, Construções e Ferrovias S.A
---------------------------------------------------------------------------
codigo_ug
      n  missing distinct     Info     Mean      Gmd      .05      .10
  18254       0     1919        1   223073   122090   120195   130058
    .25      .50      .75      .90      .95
 155009   160192   250052   393023   512006
lowest :  80003 110001 110096 110097 110099, highest: 910869 919820 925141 926015 926394
---------------------------------------------------------------------------
nome_ug
      n  missing distinct
  18254       0     1891
lowest : 1 BATALHAO DE COMUNICACOES                1 BATALHAO DE GUARDA                 1 BATALHAO DE INFANTARIA DE SELVA(AEROMOVEL)  1 BATALHÃO FERROVIARIO              1 GRUPO DE ARTILHARIA ANTIAEREA
highest: UTFPR - CAMPUS PATO BRANCO                UTFPR - CAMPUS PONTA GROSSA UTFPR - CAMPUS TOLEDO                  UTFPR CAMPUS SANTA HELENA           VALEC ENGENHARIA CONSTRUÇÕES E FERROVIAS S/A.
---------------------------------------------------------------------------
valor_inicial_compra
      n  missing distinct     Info     Mean      Gmd      .05      .10
  18254       0    13871        1  1180856  2111052     1766     4470
    .25      .50      .75      .90      .95
  19101    97487   402063  1432763  3800000
lowest :        0.00        0.01        0.03        0.07        0.10
highest: 274700000.00 276246850.32 346661548.80 364422565.39 472746000.48
---------------------------------------------------------------------------
valor_final_compra
      n  missing distinct     Info     Mean      Gmd      .05      .10
  18254       0    13918        1  1224965  2196083     1763     4439
    .25      .50      .75      .90      .95
  19200    98144   407466  1442432  3800000
lowest :        0.00        0.01        0.07        0.10        0.19
```

highest: 337350765.31 346661548.80 364422565.39 396388729.62 472746000.48

---

diferenca_compra

n	missing	distinct	Info	Mean	Gmd	.05	.10
18254	0	622	0.1	44109	101898	0	0

.25	.50	.75	.90	.95
0	0	0	0	0

lowest : -14000000 -10957023 -6366525 -2422942 -2333635

highest: 27015841 34453924 41121932 187417092 220326895

-1.4e+07 (1, 0.000), -1e+07 (1, 0.000), -6e+06 (1, 0.000), -2e+06 (11, 0.001), 0 (18181, 0.996), 2e+06 (25, 0.001), 4e+06 (9, 0.000), 6e+06 (4, 0.000), 8e+06 (5, 0.000), 1e+07 (7, 0.000), 1.4e+07 (1, 0.000), 1.8e+07 (1, 0.000), 2e+07 (1, 0.000), 2.2e+07 (1, 0.000), 2.8e+07 (1, 0.000), 3.4e+07 (1, 0.000), 4.2e+07 (1, 0.000), 1.88e+08 (1, 0.000), 2.2e+08 (1, 0.000)

For the frequency table, variable is rounded to the nearest 2000000

---

aumento_valor_contrato

n	missing	distinct
18254	0	2

Value	N	S
Frequency	17734	520
Proportion	0.972	0.028

Pela impressão das estatísticas básicas, já foi possível identificar que as seguintes variáveis possuem valores *missing*:

---

qtde_meses_entr_socio

n	missing	distinct	Info	Mean	Gmd	.05	.10
16924	1330	412	1	81.68	82.2	8	12

.25	.50	.75	.90	.95
22	52	113	197	253

lowest : 5 6 7 8 9, highest: 501 511 518 545 609

---

nro_socios

n	missing	distinct	Info	Mean	Gmd	.05	.10
16924	1330	47	0.919	2.962	2.494	1	1

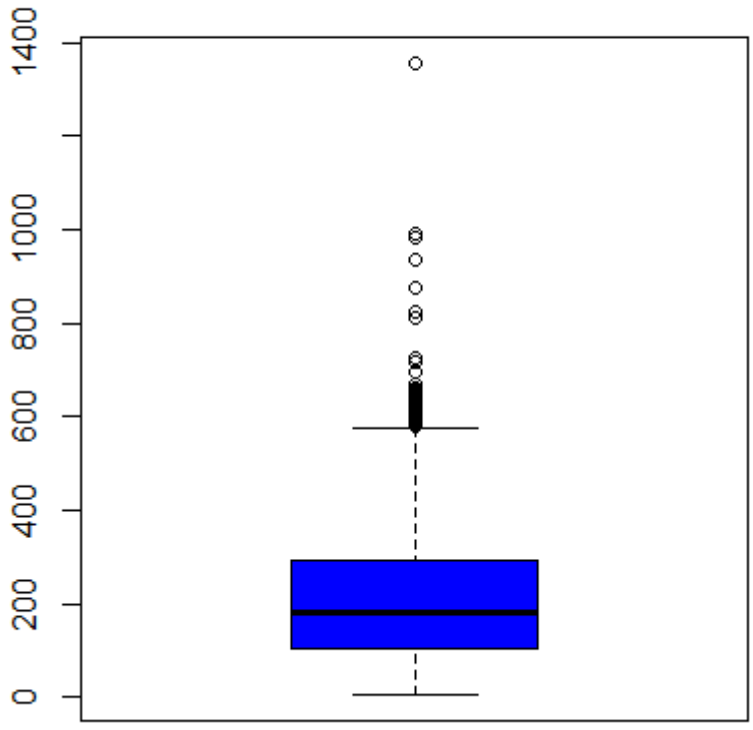
.25	.50	.75	.90	.95
1	2	3	6	7

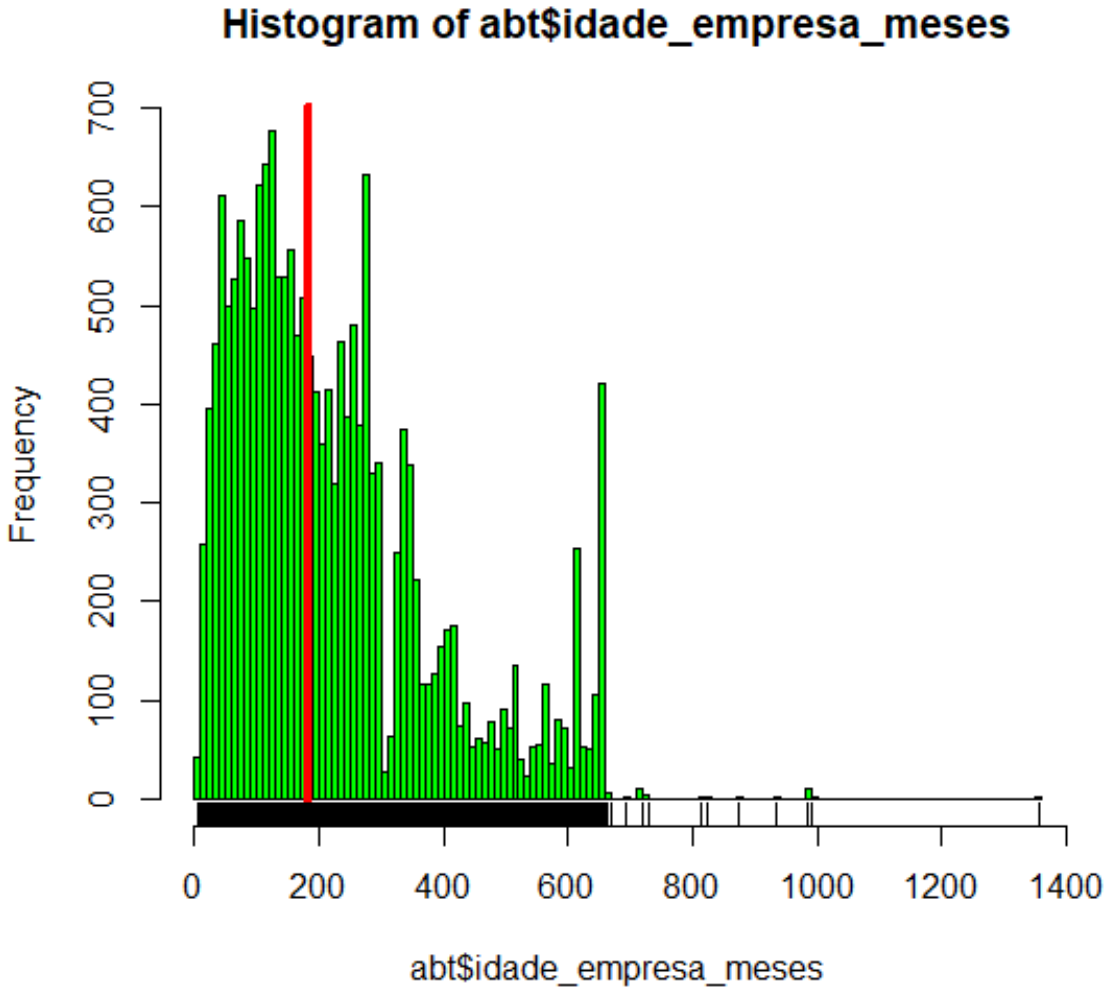
lowest : 1 2 3 4 5, highest: 166 171 181 254 362

Como a taxa de missing de ambas as variáveis em relação ao dataset global está em torno de 7,8% será feita a substituição destes valores missing pela média em cada uma delas. Como a taxa é baixa, esta medida não afetará significativamente a distribuição.

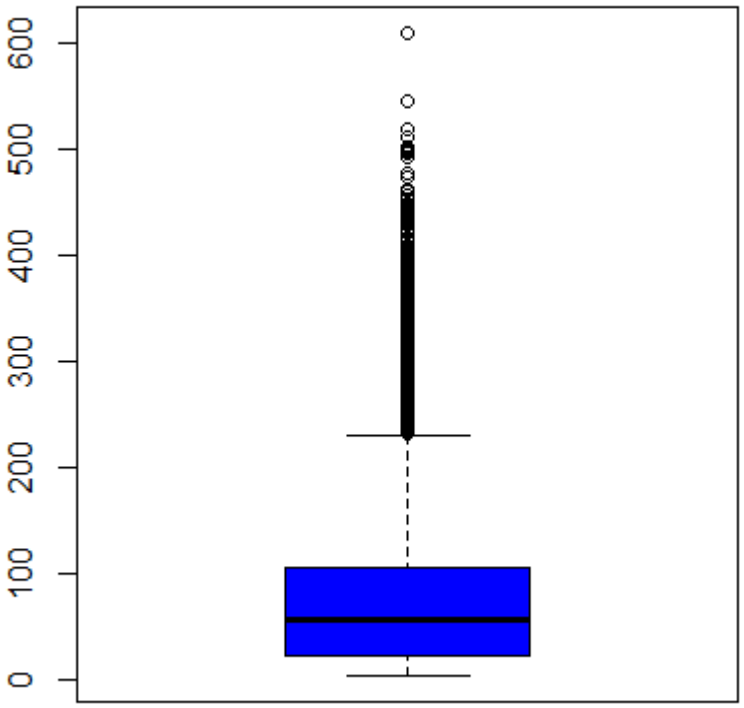
Seguem gráficos de variáveis do dataset:

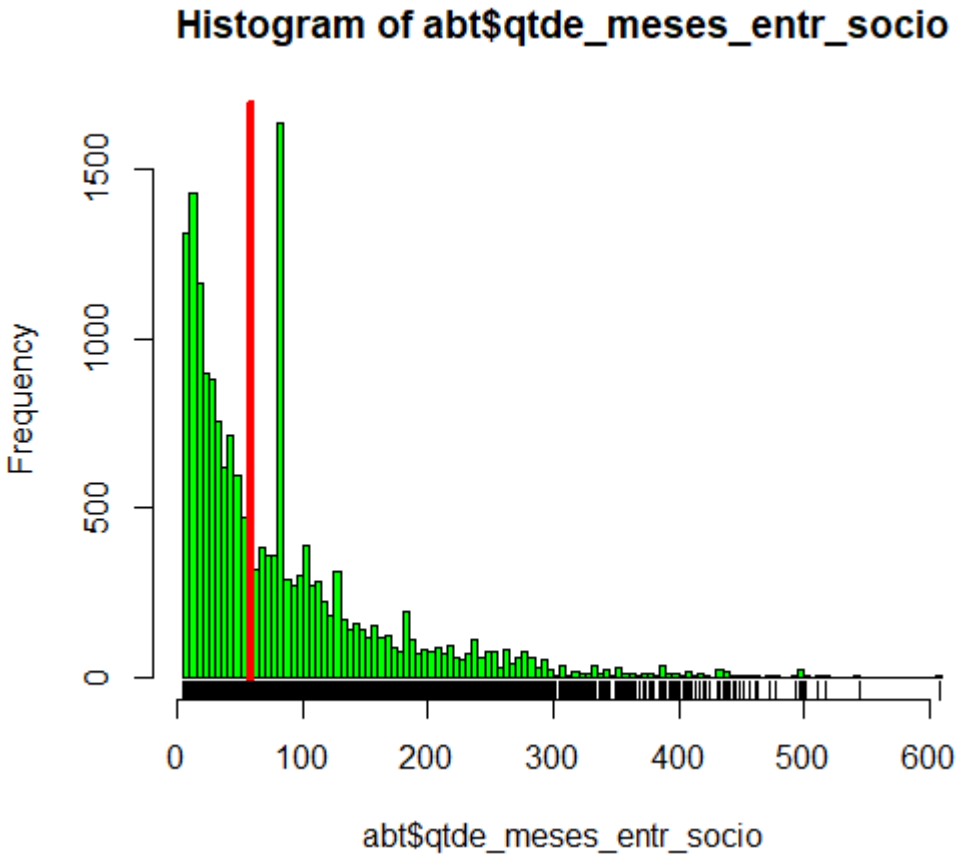
Variável: idade_empresa_meses



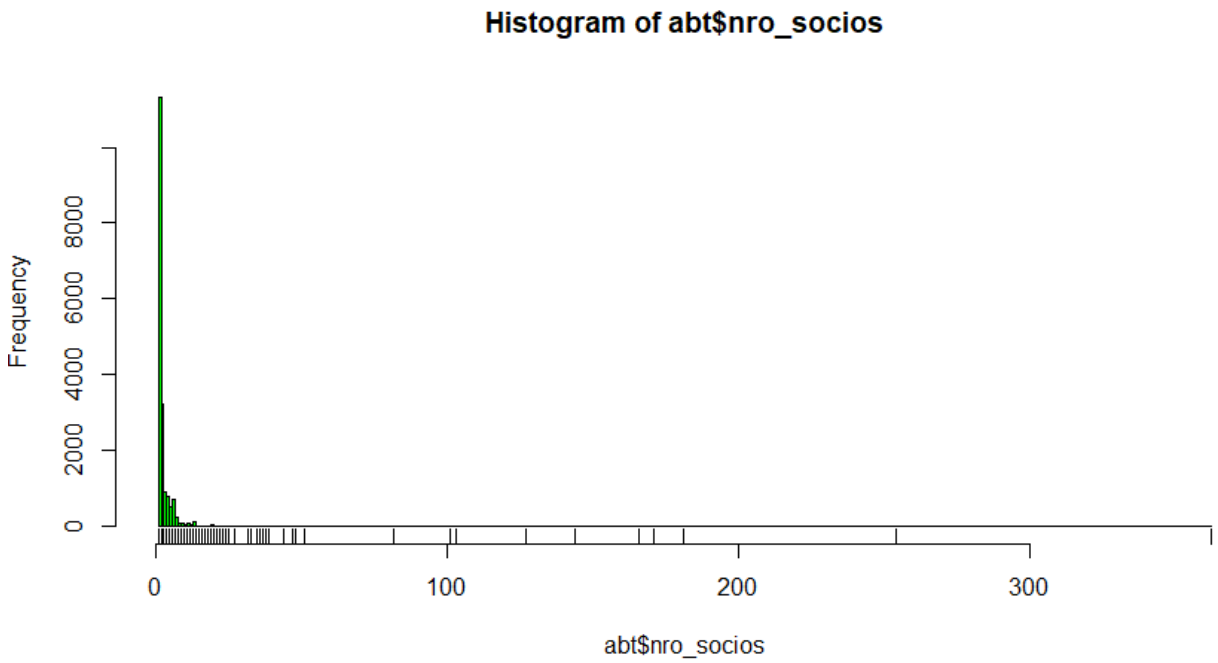

Histogram of abt$idade_empresa_meses
Frequency
0
100
200
300
400
500
600
700
0
200
400
600
800
1000
1200
1400
abt$idade_empresa_meses

Variável: qtde_meses_entr_socio

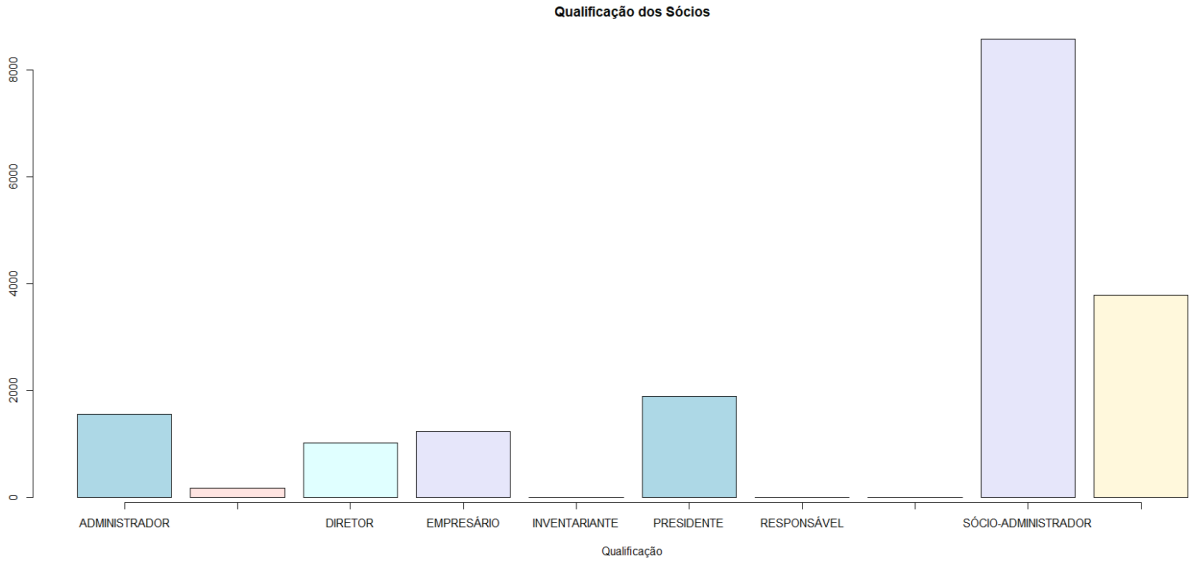




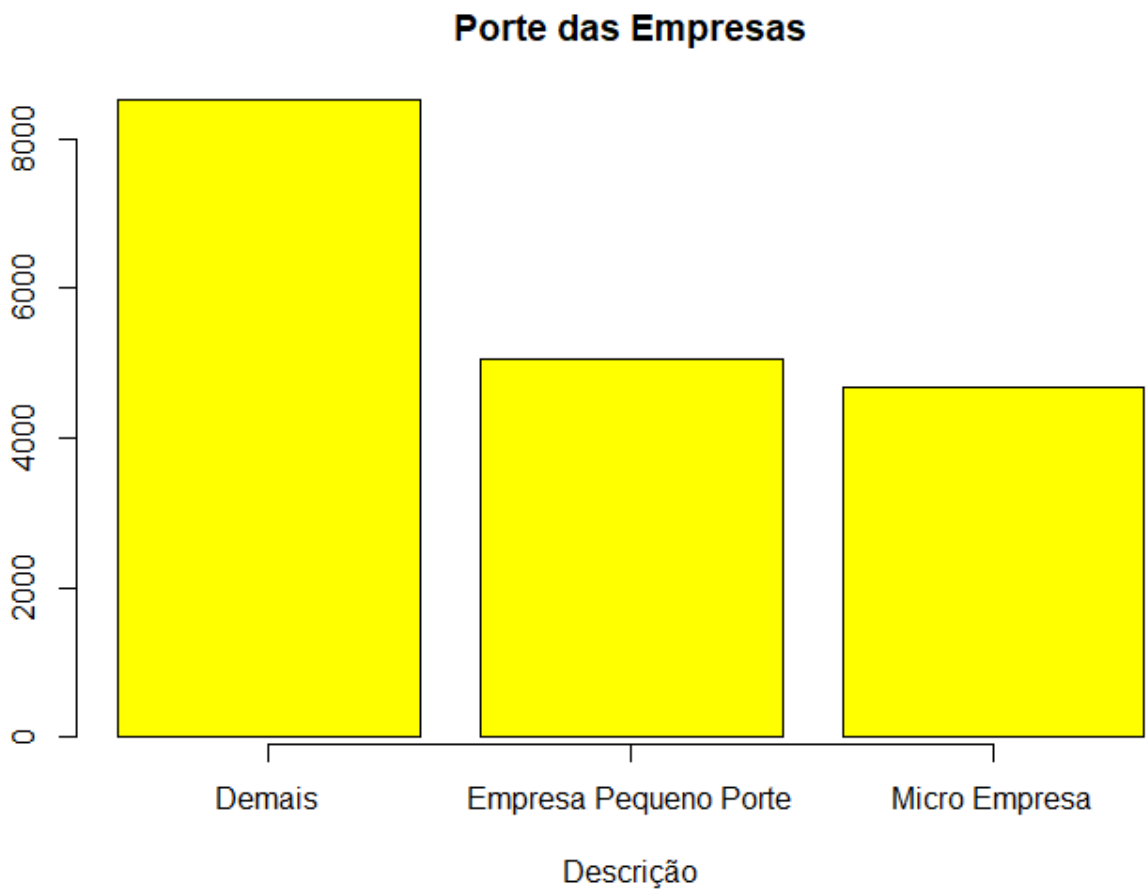
Variável: nro_socios



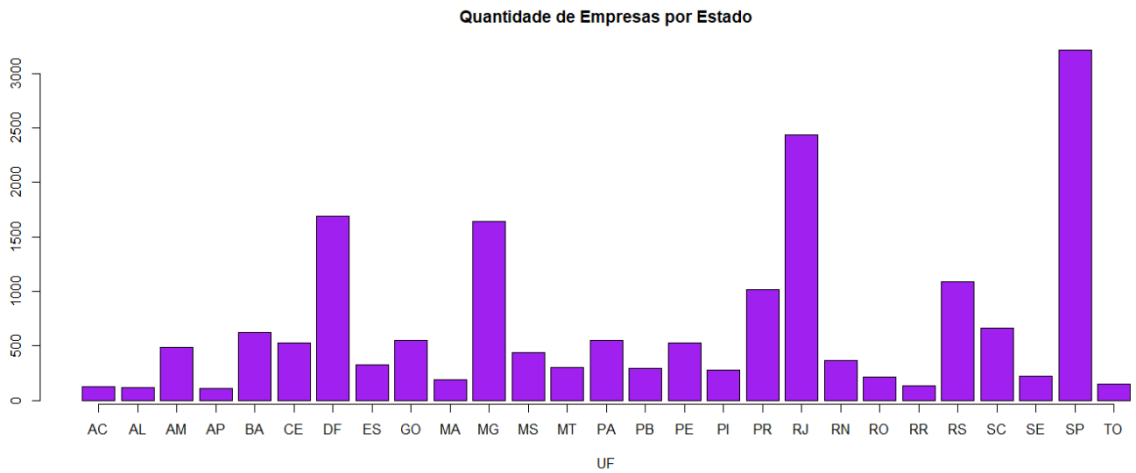
Variável: nome_qualif_socio



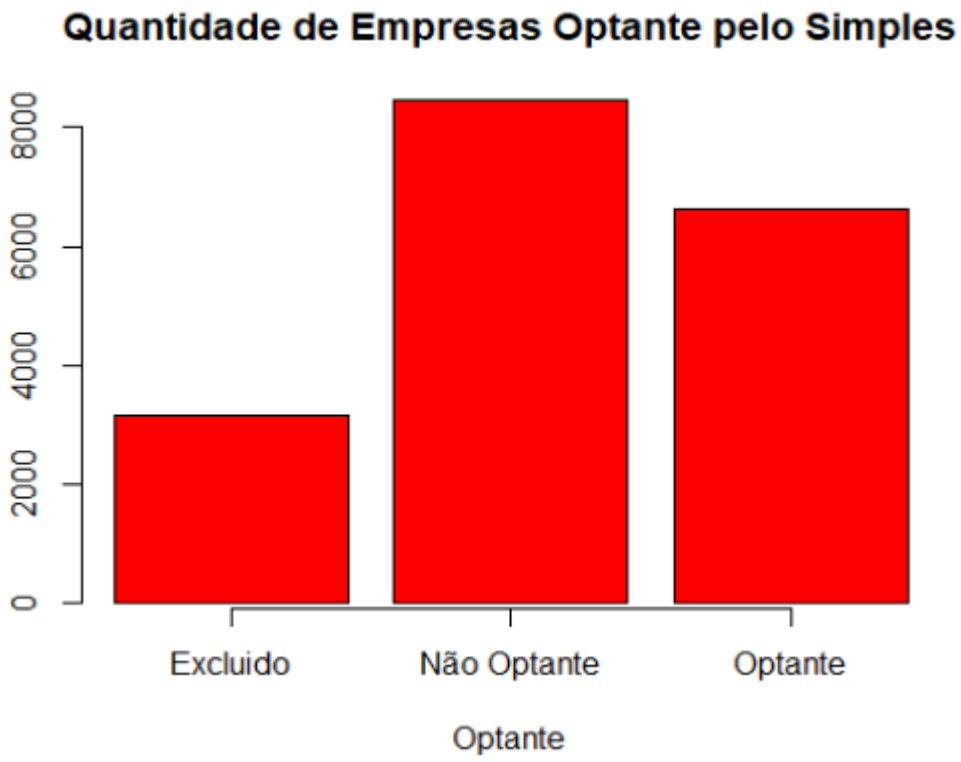
Variável: nm_porte



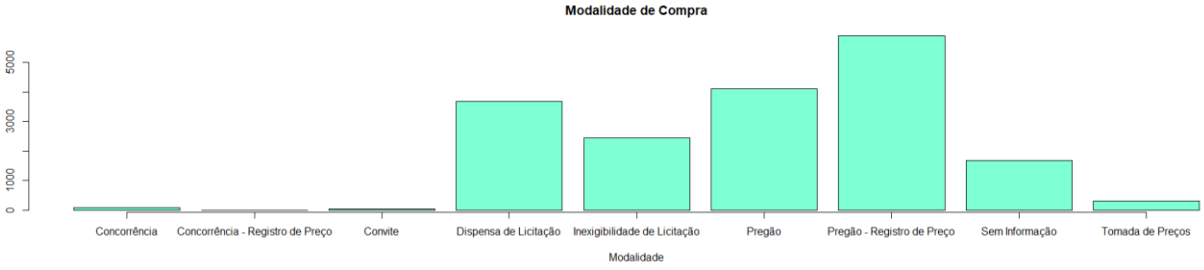
Variável: uf



Variável: nm_optante_simples



Variável: modalidade_compra



Uma parte importante na etapa de análise e exploração dos dados é a identificação das variáveis mais importantes do conjunto dos dados, ou seja, aquelas que possuem maior poder preditivo. Algumas medidas estatísticas são utilizadas para este fim, dentre elas o WOE – Weight of Evidence e IV – Information Value.

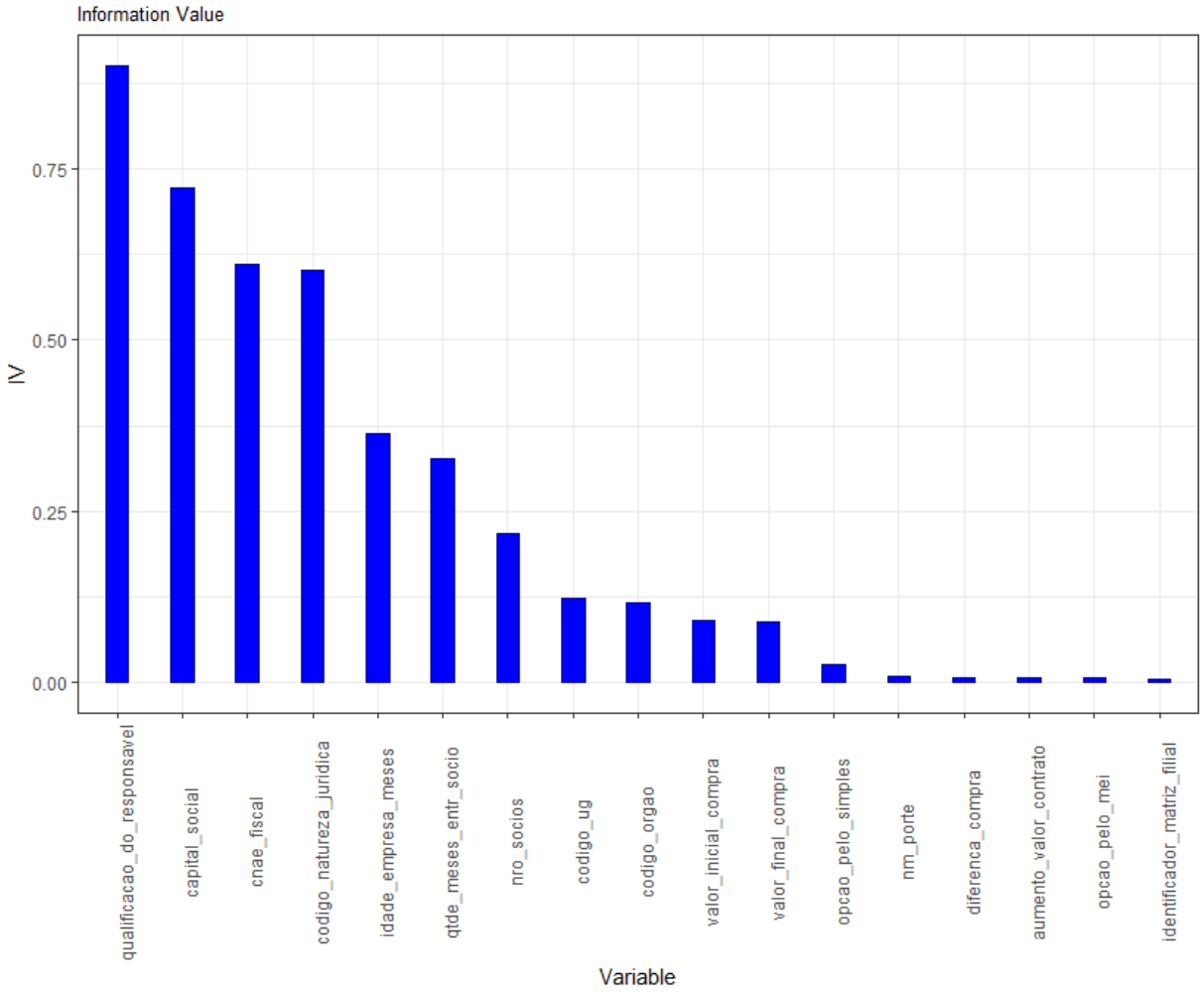
Para realizar este estudo, utilizaremos a segunda métrica IV.

A métrica IV é uma técnica muito útil para selecionar variáveis importantes dentre o grupo das variáveis explanatórias. Ou seja, nos dá a medida de quanto uma variável preditora é boa em distinguir entre uma resposta binária (por exemplo, ‘0’ ou ‘1’) em algumas variáveis resposta. Um IV muito baixo de uma variável preditora significa que ela não terá poder suficiente para classificar corretamente uma variável resposta e portanto deve ser removida como variável explicativa.

A interpretação do IV deve ser feita de acordo com a seguinte tabela:

Information Value	**Poder de Previsão**
< 0,02	Sem utilidade para previsão
0,02 – 0,1	Baixo poder preditivo
0,1 – 0,3	Médio poder preditivo
0,3 – 0,5	Forte poder preditivo
> 0.5	Muito bom ou suspeita

Realizando este estudo no dataset utilizado neste trabalho, obtemos o seguinte gráfico que mostra a capacidade de predição de cada variável considerada:



Desta forma, de acordo com a tabela da interpretação do IV, as variáveis consideradas boas preditoras são, na ordem decrescente de importância:

1. qualificacao_do_responsavel
2. capital_social
3. cnae_fiscal
4. codigo_natureza_juridica
5. idade_empresa_meses
6. qtde_meses_entr_socio
7. nro_socios

Porém, a variável qualificacao_do_responsavel apresentou um IV muito alto e de acordo com a interpretação deve ser considerada “suspeita”. Portanto, também não será considerada no modelo.

5. Criação de Modelos de Machine Learning

Após a fase de processamento, exploração e entendimento dos dados, iniciou-se a fase de construção de modelos. Para esta fase foi utilizada a Linguagem R na versão 3.6.3 em conjunto com o RStudio. R é uma linguagem de programação orientada a objetos e muito utilizada para a manipulação, análise e visualização dos dados e claro, para o desenvolvimento de modelos de machine learning.

Machine Learning é um método de análise de dados que automatiza a criação e desenvolvimento de modelos estatísticos ou analíticos utilizando algoritmos que aprendem com os dados. Este aprendizado permite que sejam encontrados padrões e insights ocultos nos dados de forma que seja possível fazer previsões ou classificações que ajudem a responder problemas de negócios.

Existem basicamente 2 tipos de modelos de aprendizagem: supervisionados e não supervisionados. Nos modelos supervisionados os dados são apresentados ao algoritmo junto com uma marcação ou rótulo mostrando já a resposta do passado. Assim os algoritmos aprendem as relações das variáveis preditoras com a variável resposta o que possibilita a realização de previsões quando o modelo for apresentado a dados novos, não rotulados. Já os modelos não supervisionados os dados são apresentados sem a variável resposta e o algoritmo tem que buscar as relações entre as observações sem essa “ajuda”. Os modelos de aprendizagem supervisionada podem ainda ser divididos em modelos de classificação ou regressão. Os modelos de classificação são aqueles em que a variável resposta é uma variável categórica ou binária, assumindo valores “0” ou “1”, “SIM” ou “NÃO” e podem conter mais de duas categorias. Neste caso busca-se classificar os dados em uma das categorias. Já os modelos de regressão possuem uma variável resposta do tipo contínua, como por exemplo, o preço de venda de um imóvel. Para prever esta variável resposta o algoritmo busca as relações entre as variáveis preditoras e a variável resposta.

Para o problema apresentado neste trabalho, optou-se por utilizar um modelo de aprendizagem supervisionada de classificação. Os dados coletados possuem uma variável resposta que permite rotular as observações e assim realizar uma aprendizagem supervisionada. Além disso, como a variável resposta é do tipo

categórica (“0” ou “1”) será realizada uma previsão de classificação em determinada categoria ou classe.

Considerando-se as definições acima expostas e baseadas nas características dos dados a serem utilizados optou-se por um modelo de aprendizagem supervisionada com algoritmo de árvore de decisão. As árvores de decisão são bastante populares por causa de seu algoritmo intuitivo. A sua saída consiste em regras que são facilmente compreensíveis pelos seres humanos, ou seja, não são considerados “caixas pretas”.

Entre as vantagens do uso das árvores de decisão estão:

1. É uma das formas mais rápidas de se identificar variáveis mais significativas e a relação entre elas. Com as árvores de decisão podemos criar novas variáveis que ajudem a predizer a variável resposta;
2. De fácil entendimento. A visualização de uma árvore torna o problema mais fácil de ser compreendido, não precisa nenhum conhecimento estatístico para ler e interpretar, alem do que a sua representação gráfica também ajuda a entender mais facilmente. Isso é muito importante na hora de apresentar as respostas aos tomadores de decisão. Fica mais fácil de explicar;
3. Menor necessidade de se limpar dados em comparação com outras técnicas mais sofisticadas que são muito sensíveis à qualidade dos dados apresentados. São menos influenciados por pontos fora da curva ou *outliers* e nem por valores *missing*. Os algoritmos conseguem lidar bem com essas questões;
4. Pode manipular tanto variáveis numéricas ou categóricas;
5. Não precisam necessariamente de transformações nos dados como normalizações ou padronizações.

Entre as desvantagens estão:

1. As árvores sofrem muito de super ajuste ou *overfitting*. Por este motivo que são utilizadas as técnicas de poda ou *prunning* de modo a tornar o modelo mais generalizável. Se ele não for genérico o suficiente só servirá para os dados aos quais foi treinado;
2. Dependendo do tamanho da árvore, pode ser de difícil entendimento para quem não tem familiaridade com a lógica de tabelas de decisões;

Mais especificamente, o algoritmo de árvore de decisão escolhido para o desenvolvimento do modelo foi o C5.0 descrito em mais detalhes no link:

https://cran.r-project.org/web/packages/C50/vignettes/C5.0.html

Para um estudo mais aprofundado das vantagens deste algoritmo sobre os anteriores ID3 e C4.5, consultar também os links:

https://www.rulequest.com/see5-info.html

https://rulequest.com/see5-comparison.html

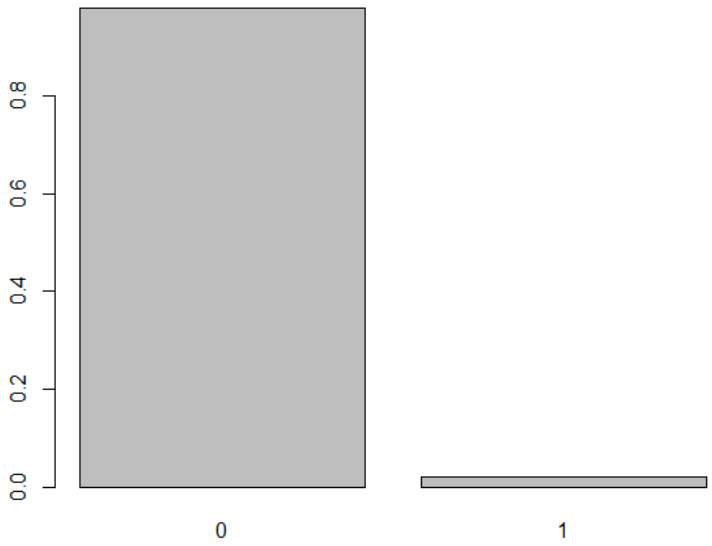
Detalhamento a seguir dos passos utilizados para a construção do modelo na linguagem R. O script R completo utilizado será disponibilizado no anexo deste trabalho.

- Etapa 1: dividindo a base em treinamento e teste

Para esta divisão, foi feito um estudo da proporção da variável resposta na base total:

```
        0         1
0.97825134 0.02174866
```

Graficamente, temos a seguinte representação:



Pode-se constatar que proporção da classe “1” que é o rótulo utilizado para as empresas inidôneas que foram impedidas de contratar é muito menor (2%) do que a classe “0” (98%). O que de certa forma é esperado para modelos de identificação de fraudes. As fraudes, em tese, são eventos não frequentes em relação aos eventos normais. Se assim não fosse, então algo estaria errado. Conclui-se que a base está completamente desbalanceada e será necessária a aplicação de alguma técnica para balancear pois o modelo pode não ter um bom desempenho nestas condições. Mas será criado um modelo com esta base desbalanceada, será verificado seu desempenho, e posteriormente a base será balanceada e será feito o comparativo entre eles.

A base foi dividida em: 70% para treinamento e 30% para teste. A proporção da variável resposta nestas duas bases manteve-se a mesma da base total:

Treinamento:

0	1
0.9797292	0.0202708

Teste:

0	1
0.97480372	0.02519628

Foi executado o treinamento do modelo na base já com a opção de poda ativada e obteve-se o seguinte resultado:

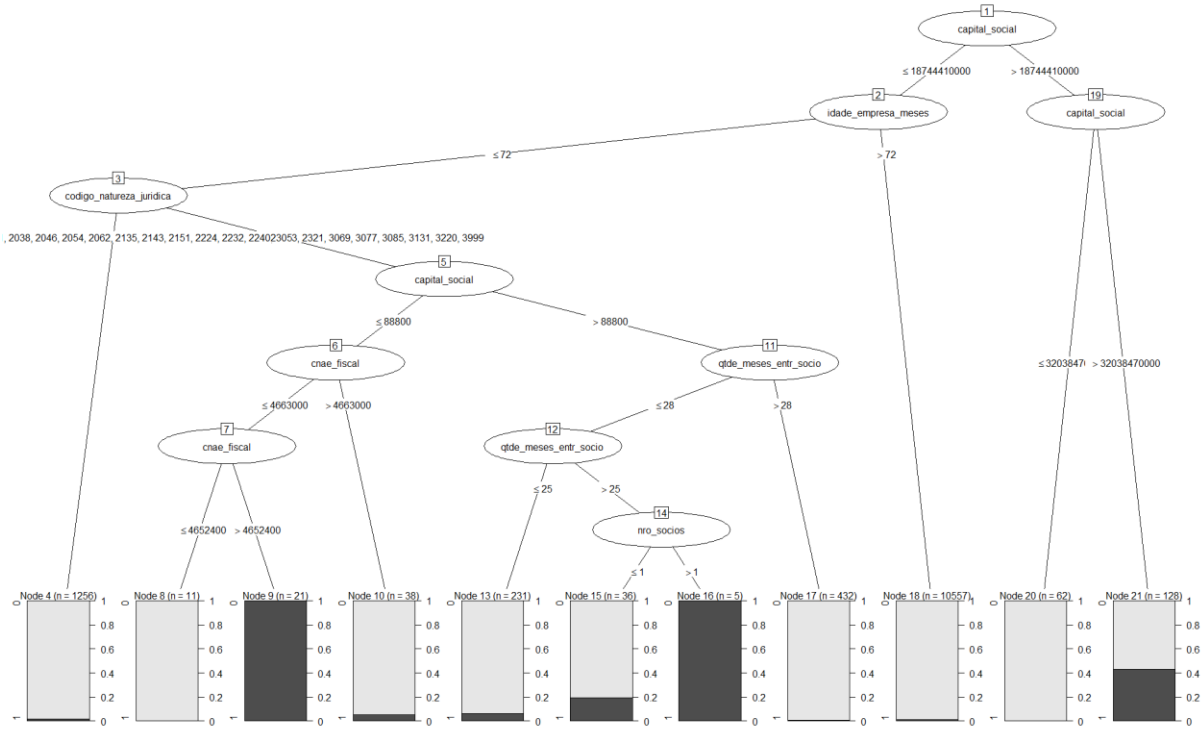
```
Call:
C5.0.formula(formula = inidonea ~ codigo_natureza_juridica
 abt_treino, method = "class", control
 = C5.0Control(noGlobalPruning = FALSE, minCases = 5))

Classification Tree
Number of samples: 12777
Number of predictors: 6

Tree size: 11

Non-standard options: attempt to group attributes,
 minimum number of cases: 5
```

Foi gerado um total de 11 regras (folhas). A seguir o gráfico da árvore gerada:



Detalhes do modelo gerado:

```
Evaluation on training data (12777 cases):

           Decision Tree
         ----------------
         Size      Errors

           11  178( 1.4%)   <<


          (a)   (b)    <-classified as
         ----  ----
        12518          (a): class 0
          178    81    (b): class 1


        Attribute usage:

        100.00% capital_social
         99.00% idade_empresa_meses
         15.89% codigo_natureza_juridica
          5.51% qtde_meses_entr_socio
          0.55% cnae_fiscal
          0.32% nro_socios
```

Pode-se ver a matriz de confusão gerada e também a contribuição de cada variável para o modelo. Abaixo, outra informação da importância de cada variável no modelo:

```
                          Overall
capital_social             100.00
idade_empresa_meses         99.00
codigo_natureza_juridica    15.89
qtde_meses_entr_socio        5.51
cnae_fiscal                  0.55
nro_socios                   0.32
```

Desta forma observa-se um melhor entendimento de quais variáveis mais contribuíram para o modelo.

Próximo passo é verificar a acurácia do modelo e ver como está o seu desempenho quando aplicamos o modelo na base de teste. Ao executar a predição na base de testes obtém-se o seguinte resultado:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5339    0
         1   97   41

               Accuracy : 0.9823
                 95% CI : (0.9784, 0.9856)
    No Information Rate : 0.9925
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4518

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.000000
            Specificity : 0.982156
         Pos Pred Value : 0.297101
         Neg Pred Value : 1.000000
             Prevalence : 0.007486
         Detection Rate : 0.007486
   Detection Prevalence : 0.025196
      Balanced Accuracy : 0.991078

       'Positive' Class : 1
```

Por meio da matriz de confusão obtém-se alguns indicadores de desempenho dos modelos de classificação dentre os quais a acurácia. É uma tabela que indica os erros e acertos de um modelo comparando com o resultado esperado que seriam os rótulos das classes. A imagem abaixo mostra um exemplo de uma matriz de confusão:

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

- Verdadeiros Positivos: classificação correta da classe Positivo;
- Falsos Negativos: erro em que o modelo previu a classe Negativo quando o valor real era classe Positivo;
- Falsos Positivos: erro em que o modelo previu a classe Positivo quando o valor real era a classe Negativo;

- Verdadeiros Negativos: classificação correta da classe Negativo.

A Acurácia indica uma performance geral do modelo, ou seja, dentre todas as classificações quantas o modelo classificou corretamente.

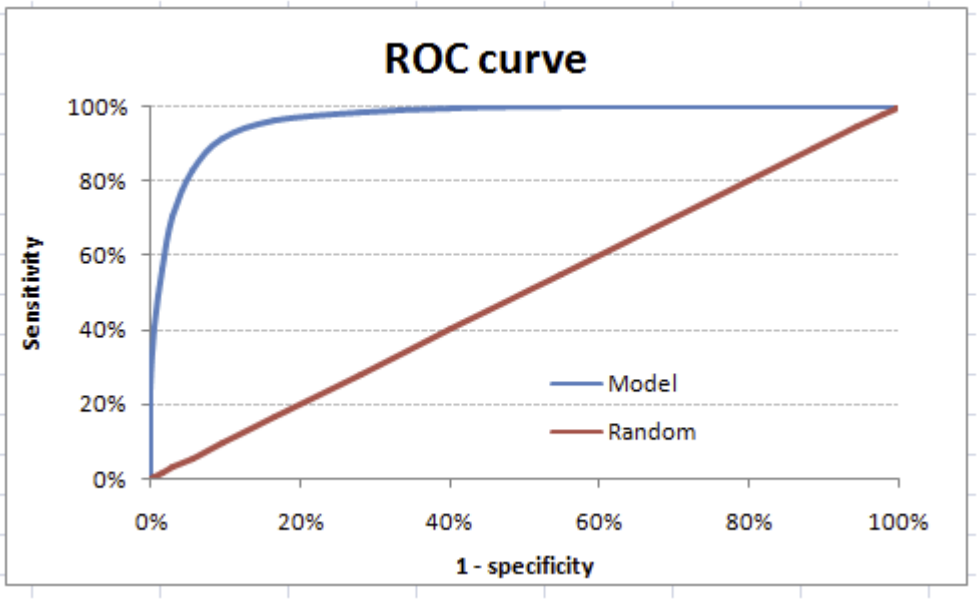
Cálculo da Acurácia:

$$Acurácia = \frac{Verdadeiros\ Positivos\ (TP) + Verdadeiros\ Negativos\ (VN)}{Total}$$

A Acurácia do modelo gerado e aplicado na base de testes foi de 98,32%. Mas este desempenho foi em uma base desbalanceada e portanto pode não ser considerado um bom desempenho.

Deve-se considerar uma outra medida de avaliação e neste caso podemos utilizar a curva ROC e verificar a área abaixo da curva.

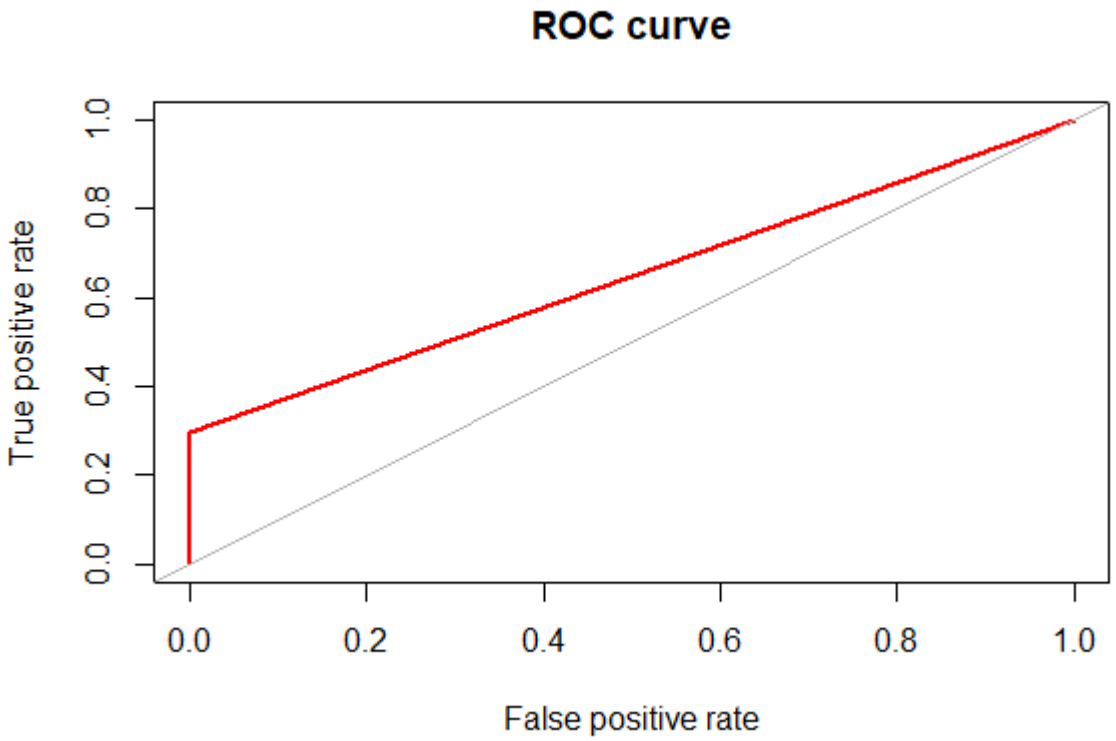
Receiver Operating Characteristics (ROC) e Area Under the Curve (AUC) são duas métricas muito utilizadas para avaliação de modelos de classificação. O cálculo dessas métricas é muito semelhante à matriz de confusão. Neste caso, plota-se a Sensibilidade, que é a taxa de Verdadeiro Positivo, e (1-Especificidade), que é a taxa de Falso Positivo. Passa-se também um parâmetro de corte, um Threshold, onde geralmente adota-se o valor de 0,5. Abaixo gráfico de uma curva ROC:



A interpretação é que quanto mais alto e mais distante da linha diagonal a curva estiver, melhor. Pode-se também transformar esta área no gráfico em um número que é o AUC. A AUC é um número de 0 a 1 que mostra como está o

desempenho do modelo utilizando como cálculo a Taxa de Falso Positivo, a Taxa de Verdadeiro Positivo e o Threshold definido. Neste caso, a interpretação é que quanto mais próximo de 1 este número estiver, melhor.

Então calculando-se a ROC e AUC para o modelo de árvore de decisão, obtém-se:



O indicador AUC foi calculado em:

Area under the curve (AUC): 0.649

A próxima etapa agora será realizar o balanceamento da base. Bases desbalanceadas são um problema para algoritmos de machine learning pois eles podem ter dificuldade em aprender os padrões relacionados com a classe de menor representatividade, como é o caso da presente base de dados. Assim, usando-se esta base desbalanceada, o algoritmo tende a gerar um modelo que favoreça a classificação dos novos dados apresentados na classe majoritária, o que leva a resultados inconsistentes ainda mais em se tratando de modelos de detecção de fraudes.

Algumas técnicas são usadas para lidar com esse problema, dentre elas a redefinição do tamanho do conjunto de dados por meio de:

1. Undersampling: diminui a proporção de observações da classe majoritária
2. Oversampling: aumenta a proporção de observações da classe minoritária

Será utilizado neste trabalho a técnica de Oversampling conhecida como SMOTE – Synthetic Minority Oversampling Technique. Na técnica clássica de Oversampling, os dados da classe minoritária são duplicados apenas. Ele aumenta a proporção mas não traz nenhuma nova informação ou variação que ajude o modelo de machine learning. Já o SMOTE, em linhas gerais, utiliza um algoritmo chamado K-nearest neighbor para criar dados sintéticos onde dados são escolhidos randomicamente da classe minoritária e depois vizinhos próximos a estes são escolhidos. Assim, novos dados são gerados usando-se os dados escolhidos randomicamente e os vizinhos também escolhidos randomicamente.

Aplicando-se esta técnica nas bases de treinamento e testes, pode-se observar a mudança na proporção das classes:

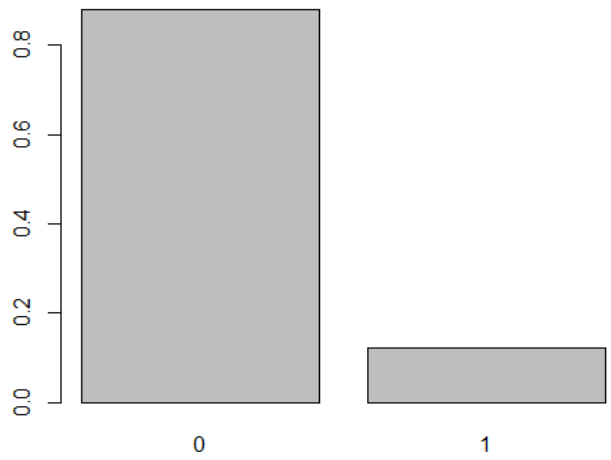
Treinamento:

```
        0         1
0.8791209 0.1208791
```

Teste:

```
        0         1
0.8791209 0.1208791
```

Verificando graficamente a proporção:



Executando o modelo de aprendizagem na nova base balanceada os resultados modificaram bastante, como pode-se observar:

```
Call:
C5.0.formula(formula = inidonea ~ codigo_natureza_juridica
 = abt_treino_smote, method = "class", control
 = C5.0Control(noGlobalPruning = FALSE, minCases = 5))

Classification Tree
Number of samples: 23569
Number of predictors: 6

Tree size: 110

Non-standard options: attempt to group attributes,
 minimum number of cases: 5
```

Aumentou muito a quantidade de folhas geradas. Observando-se agora a contribuição das variáveis, também nota-se uma acentuada modificação:

	Overall
capital_social	100.00
nro_socios	96.95
codigo_natureza_juridica	88.04
cnae_fiscal	54.37
qtde_meses_entr_socio	21.42
idade_empresa_meses	16.59

Verificando-se a Acurácia do modelo na base de teste balanceada, nota-se uma mudança marginal, de 98,23% para 97,56%.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 10986    54
         1   252  1266

               Accuracy : 0.9756
                 95% CI : (0.9728, 0.9783)
    No Information Rate : 0.8949
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8785

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9776
            Specificity : 0.9591
         Pos Pred Value : 0.9951
         Neg Pred Value : 0.8340
             Prevalence : 0.8949
         Detection Rate : 0.8748
   Detection Prevalence : 0.8791
      Balanced Accuracy : 0.9683

       'Positive' Class : 0
```

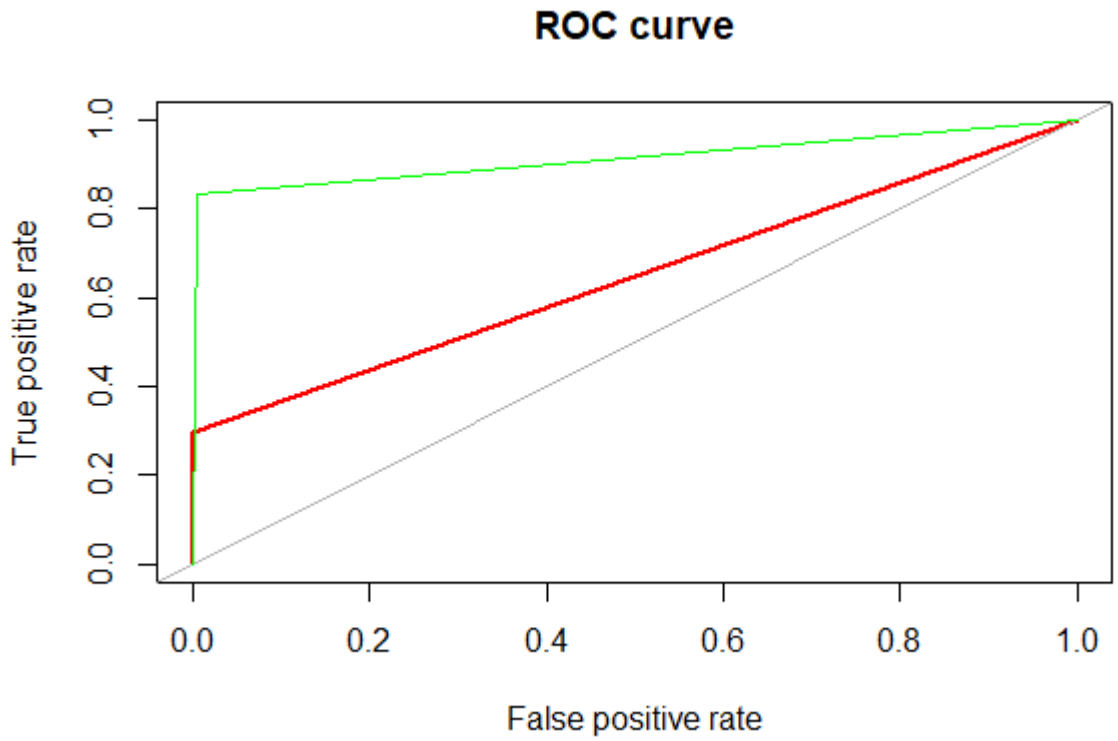
Com relação aos indicadores de ROC e AUC:

A linha verde abaixo representa a curva ROC para o novo modelo gerado na base balanceada. Já se observa a melhora de desempenho deste novo modelo.

Resultado do AUC:

Area under the curve (AUC): 0.915

O AUC subiu consideravelmente de 0,649 para 0,918 o que confirma que este modelo treinado e testado nas novas bases balanceadas tem melhor desempenho.



6. Apresentação dos Resultados

Os problemas relacionados a fraudes, seja em organizações privadas ou públicas, ficam cada vez mais sofisticados e mais frequentes. Com a crescente digitalização e com cada vez mais dispositivos pessoais que capturam informações incessantemente o ambiente fica propício às mais diversas formas de fraudes.

A questão se agrava quando falamos em fraudes que ocorrem em organizações públicas ou em governos pois o que está em jogo neste caso é a dilapidação do erário causando perdas significativas para a sociedade e prejudicando a implementação de políticas públicas para os cidadãos. Isso sem levar em conta os prejuízos que são causados na economia de um país tais como distorções e ineficiências que afetam a competitividade de um país.

Por estas razões reforça-se a importância da transparência principalmente no esforço em se tornar públicas as informações que são geradas por governos e no caso do presente trabalho, a disponibilização pública de licitações e contratos para que se possam realizar atividades de controle. Está claro que estes controles já existem e são muito bem conduzidos pelos órgãos de controle do Estado. E assim a Ciência de Dados pode contribuir de maneira efetiva com estes órgãos ajudando a identificar, caracterizar e realizar previsões.

Desta forma, o que este trabalho se propõe é ajudar a criar mecanismos, no caso modelos estatísticos, que ajudem esses órgãos de controle ou qualquer pessoa ou organização a ajudar a minimizar os efeitos causados pelas fraudes. E para isso é preciso alguma ferramenta que ajude na sua identificação.

Seguindo a metodologia proposta pela direção do curso, será utilizado o modelo de Data Science Workflow Canvas desenvolvido por Vasandani.

O presente trabalho desenvolve uma proposta de modelagem de fraudes em compras da Administração Pública Federal. Este é o problema a ser resolvido, identificar e prever possíveis fraudes nas compras e contratos firmados.

Um dos produtos (saídas) esperados deste trabalho é o desenvolvimento de um modelo preditivo de classificação de possíveis fraudes que ocorrem em contratos governamentais. Este modelo deve ter um desempenho satisfatório. Para esta modelagem existe uma lista de empresas que foram investigadas pelos órgãos de controle e foram consideradas inidôneas. Esta lista é pública e pode ser consultada por qualquer cidadão. Esta é a nossa variável resposta. Para variáveis explicativas foram consideradas bases de dados públicos de várias fontes.

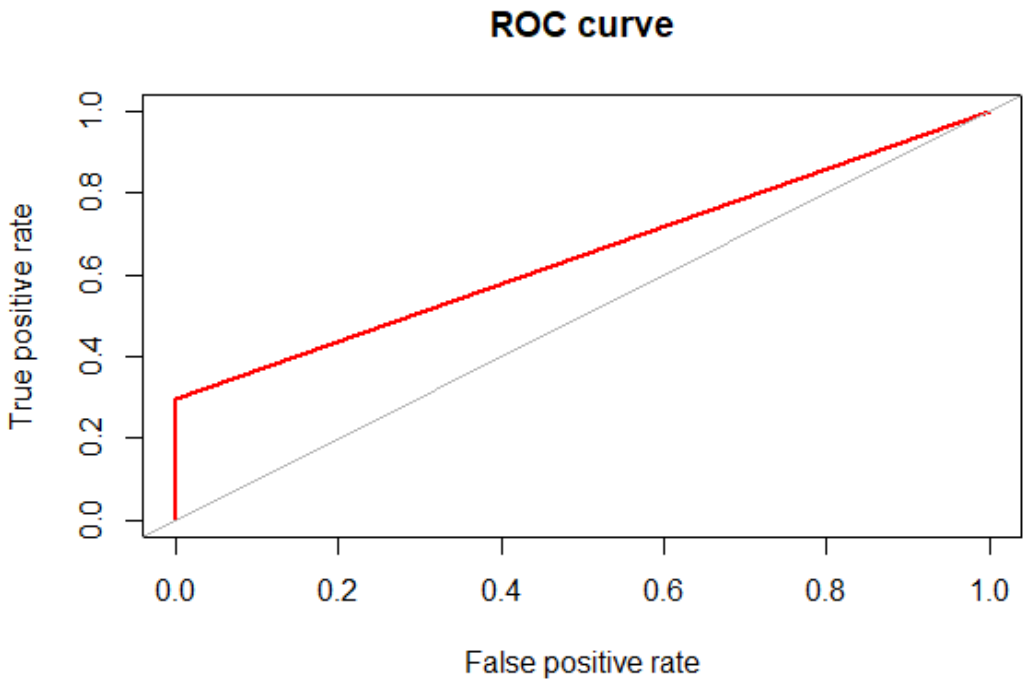
Para aquisição dos dados foram buscadas fontes públicas de dados, tais como os portais de transparência e o portal de dados abertos do Governo Federal. Em

princípio existem dados suficientes que ajudaram no desenvolvimento do modelo. Porém acredito que os órgãos de controle possam ter acesso a dados mais completos e que permitam ser adicionados à lista de variáveis explicativas. Isso permitiria o desenvolvimento de modelos mais específicos e certamente melhoraria o desempenho geral. Nesta etapa foi feito um estudo para identificar quais as variáveis preditoras que oferecem um melhor poder de previsão ao modelo. Este estudo baseou-se na métrica Information Value.

Dado que a variável resposta para este problema é categórica, ou seja, se a empresa tem possibilidade de ser inidônea ou não, optou-se por um algoritmo de aprendizagem supervisionada e que fizesse previsões de classificação. Foi escolhido modelo baseado em árvore de decisão. Além de ser um modelo que possui algumas vantagens técnicas que já foram anteriormente abordadas, também é um modelo de fácil explicação e entendimento por pessoas que não têm conhecimento técnico aprofundado.

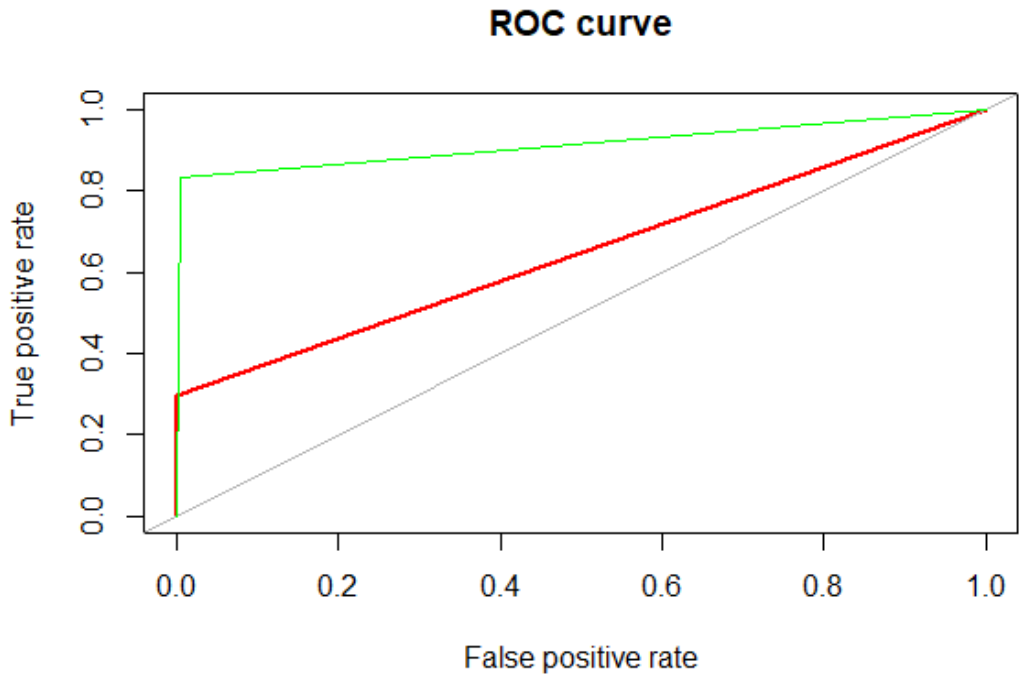
Para a avaliação do modelo foi preciso abordar e resolver um problema comum em desenvolvimento de modelos de detecção de fraudes. A pouca disponibilidade de dados rotulados como ‘negativos’. Essa é uma característica desse tipo de modelagem já que o que se espera é que a classe ‘negativa’ seja menor do que a classe ‘positiva’ porque de outra forma estaríamos convivendo apenas com fraudes. Este cenário causa um problema chamado de bases desbalanceadas. Para resolver este problema, primeiro foi desenvolvido um modelo preditivo antes do balanceamento das bases e obteve-se os seus resultados de desempenho em uma base de testes. Em seguida foi aplicada uma técnica para balancear as bases e novamente foi medido o seu desempenho. Ao se comparar os dois modelos observou-se que o modelo com bases balanceadas foi o que obteve melhor desempenho e foi este o escolhido.

Trazendo a comparação dos modelos em gráficos:



O gráfico acima mostra o desempenho do modelo desenvolvido nas bases desbalanceadas. A linha vermelha mostra esse resultado. Ela deveria estar o mais afastada possível da linha diagonal em cinza.

Agora comparando com o gráfico que mostra o desempenho do modelo gerado com as bases balanceadas:



A linha verde mostra o desempenho desse novo modelo. Pode-se ver com facilidade que o segundo modelo é melhor que o primeiro.

Para que este modelo seja efetivamente utilizado, novas bases de dados serão necessárias. Ou seja, mais dados deverão ser obtidos junto às fontes utilizadas, depois deverão ser novamente preparados e então apresentados ao modelo. A partir da aplicação do modelo em novas observações as mesmas serão classificadas (previstas) e serão marcadas na base pelo modelo as que têm indícios de serem consideradas inidôneas. Mas para isso não basta que o modelo aponte. É preciso que essa informação chegue aos órgãos responsáveis pela fiscalização e que então se façam os procedimentos necessários a fim de efetivamente validar a indicação do modelo, ou seja, confirmar se é ou não uma empresa inidônea. E essas novas marcações retroalimentam os novos modelos pois já foram classificadas e servirão para o aprendizado de novos modelos.

7. Links

Link para o vídeo: https://youtu.be/_OUsplGCwfM

Link para o repositório: https://github.com/rfartur/TCC_PUCMinas

REFERÊNCIAS

Links pesquisados durante o desenvolvimento do trabalho:

https://medium.com/@vitorborbarodrigues/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c
https://paulovasconcellos.com.br/como-saber-se-seu-modelo-de-machine-learning-est%C3%A1-funcionando-mesmo-a5892f6468b
https://cran.r-project.org/web/packages/C50/vignettes/C5.0.html
https://www.rulequest.com/see5-info.html
https://rulequest.com/see5-comparison.html
https://cran.r-project.org/web/packages/Information/index.html
https://www.r-bloggers.com/2017/08/woe-and-iv-variable-screening-with-information-in-r/
https://towardsdatascience.com/5-smote-techniques-for-oversampling-your-imbalance-data-b8155bdbe2b5
https://towardsdatascience.com/churn-analysis-information-value-and-weight-of-evidence-6a35db8b9ec5
https://pt.wikipedia.org/wiki/An%C3%A1lise_explorat%C3%B3ria_de_dados
https://www.techedgegroup.com/pt/blog/processo-de-ciencia-de-dados-exploracao-de-dados
https://www.datascience-pm.com/crisp-dm-2/
https://towardsdatascience.com/a-data-science-workflow-canvas-to-kickstart-your-projects-db62556be4d0