AGAMOTTO
TV Gazeta

inteli
instituto
de tecnologia
e liderança

Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
08/08/2022	Antonio Nassar Arthur Prado Carolina Fricks Eduardo Porto Gabriela Matias Mateus Rafael Livia Bonotto	1.1	Criação do documento Começamos as seções: 2.1, 2.2, 3.1, 4.1.1 a, 4.1.1 b, 4.1.2, 4.1.3 d, 4.1.4,
26/08/2022		1.2	Feature engineering

Sumário

1. Introdução

Informação, entretenimento e prestação de serviços de comunicação focado no estado do Espírito Santo. A Rede Gazeta é o maior grupo de comunicação capixaba, possuindo oito estações de rádio e quatro emissoras de TV aberta afiliadas à Rede Globo.

A Rede Gazeta é uma empresa de grande importância para todo o estado do Espírito Santo, pois é a maior emissora de TV aberta local contando com mais de 500 funcionários.

O problema enfrentado pela emissora é poder criar novos programas de TV de forma assertiva com decisões baseadas em dados.

2. Objetivos e Justificativa

2.1. Objetivos

O objetivo geral do parceiro é prever a audiência em um período específico de tempo e, com base nessa métrica, prever o investimento que ele terá que fazer para o programa atingir a expectativa de telespectadores, aumentar a audiência e o tempo de permanência. Também há o objetivo de estimar qual programa passar em determinado horário para obter melhores resultados.

2.2. Justificativa

Pensando no problema, surge a necessidade de criar um modelo preditivo responsável por garantir uma sequência de informações que garanta uma visão prévia dos fatores e elementos que impactam no resultado final da audiência. Desse modo, com o modelo levantado por meio dos dados listados, podemos encontrar uma forma de prever como cada fator impacta no negócio e, assim, determinar um possível crescimento da audiência, tendo como base o horário (alcance e/ou tempo de permanência), fornecendo ao parceiro o score de audiência a partir do peso de cada variável na definição do resultado - dia; tempo (por hora); audiência; share; tempo de permanência; alcance e gênero do produto.

3. Metodologia

Descreva as etapas metodológicas que foram utilizadas para o desenvolvimento, citando o referencial teórico. Você deve apenas enunciar os métodos, sem dizer ainda como ele foi aplicado e quais resultados obtidos.

3.1. CRISP-DM

CRISP-DM, abreviação para Cross Industry Standard Process for Data Mining, é uma metodologia de planejamento para mineração de dados onde, por meio de um ciclo de etapas, é possível compreender o andamento/fluxo de um processo ou análise de dados de um projeto. As etapas consistem em fases, onde após o compreendimento do modelo de negócios, se estabelece a maneira como dados são coletados e analisados. Após isso, estes dados são preparados para serem implementados e modelados conforme a necessidade de seu determinado uso. Por fim, as chamadas Instâncias dos Processos são a fase final onde esses dados são “sólidos” e assim estão prontos para serem utilizados.

3.2. Ferramentas

Descreva brevemente as ferramentas utilizadas e seus papéis (Google Collaboratory)

3.3. Principais técnicas empregadas

Descreva brevemente as principais técnicas empregadas, algoritmos e seus benefícios

4. Desenvolvimento e Resultados

De maneira geral, você deve descrever nesta seção a aplicação dos métodos aprendidos e os resultados obtidos por seu grupo em seu projeto

4.1. Compreensão do Problema

4.1.1. Contexto da indústria

a) 5 Forças de Porter

i) Principais Players

De acordo com a tabela indicada abaixo relativa aos principais players do mercado em concorrência, temos a TV Vitória (representante da Record) e TV Tribuna (representante do SBT). Porém, tendo como base os dados de audiência, foi possível entender que, apesar da força dos concorrentes dentro do mercado, atualmente o poder de alcance da audiência do parceiro é muito mais amplo. Com exceções muito pontuais de horários nos quais existe uma queda, mas o modelo preditivo busca, justamente, fornecer uma solução para que o parceiro possa ter um bom desempenho também nos horários de queda.

	TV Vitória (REC)	TV Tribuna (SBT)
Descrição	"A TV Vitória possui alcance por todos 78 municípios do Espírito Santo. Com um único sinal, é transmitida uma cobertura completa com jornalismo, entretenimento, esporte, dramaturgia e coberturas nacionais e internacionais de qualidade. Mesmo com um sinal unificado, a TV Vitória consegue, com maestria, direcionar seus conteúdos a comunidades locais."	"TV Tribuna é uma emissora de televisão brasileira sediada em Vitória, capital do estado do Espírito Santo. Opera no canal 7 (42 UHF digital) e é afiliada ao SBT. Seus estúdios localizam-se no Edifício João Santos Júnior, onde funciona toda a holding da Rede Tribuna, no bairro Ilha de Santa Maria, e sua antena de transmissão está no topo do Morro da Fonte Grande."
Pontos Fortes	- "A TV Vitória recebeu o título de melhor TV regional do Brasil por sete vezes, eleita pela Academia Brasileira de Marketing"	- "A TV Tribuna conquistou o respeito e a audiência dos telespectadores e do mercado anunciante, consolidando-se como a 2ª maior emissora de TV do Espírito Santo. A grade regional é formada por 10 programas líderes em seus segmentos, com destaque para o Tribuna Notícias 1ª Edição, que se tornou referência nacional entre as afiliadas do SBT."
Pontos Fracos	- Abaixo da faixa padrão de audiêcia da TV Gazeta (Globo)	- Abaixo da faixa padrão de audiêcia da TV Gazeta (Globo)

ii) Relação com Clientes:

O poder de barganha dos compradores diz respeito à capacidade de barganha dos clientes para com as empresas do setor. No nosso caso, essa força é uma das mais atuantes pois a audiência diz respeito ao poder de escolha dos compradores/clientes em selecionar para qual rede de televisão ou canal vão dedicar sua atenção em determinado dia ou horário. Nosso produto irá determinar justamente os parâmetros que influenciam no poder de aderência dos compradores em relação a emissora e como isso pode influenciar no negócio.

iii) Fornecedores:

No caso do mercado televisivo não existe necessariamente uma grande concorrência entre fornecedores, pois se trata de fornecedores pulverizados. Logo, existem várias opções, como por exemplo: Atores, Repórteres, Fornecedores de Equipamentos, entre outros profissionais ou empresas que atuam com o fornecimento de equipamentos e pessoas para atuarem no mercado.

iv) Novos Entrantes:

Para o mercado televisivo existe uma grande dificuldade para a atuação e implementação de novos entrantes, visto que é um mercado extremamente bem consolidado e regulado para emissoras tradicionais. Logo, novos entrantes não se apresentam como uma ameaça para o negócio pois atualmente existem diversos aspectos e fatores, até mesmo burocráticos, que limitam a criação de novos canais televisivos. Desse modo, os principais players são grandes concorrentes, mas existe pouca probabilidade de novos entrantes que podem ser inseridos no mercado.

v) Tendências do Mercado:

Pensando que atualmente existe uma grande demanda em relação ao acesso a conteúdos por meio da Internet ou *Smartphones*, a indústria televisiva como um todo vem sofrendo uma queda. “De acordo com a Agência Nacional de Telecomunicações (Anatel), a queda de assinantes de TV paga tem sido impulsionada pela mudança no comportamento dos telespectadores, que estão optando por acompanhar filmes e séries em plataformas de streaming, como Netflix e Amazon, pois oferecem conteúdos originais e serviços com um custo menor aos usuários.” [DIGILAB, 2019]. Porém, tendo como base a própria análise dos dados de audiência, temos que os serviços como *Streaming,* por exemplo, possuem um nível de audiência individual, sendo que existem os players A, B ou C, cada um deles tem sua audiência determinada individualmente, o que em comparação com a rede televisiva segue sendo um nível abaixo da audiência do parceiro.

4.1.2. Análise SWOT

Posicione aqui sua análise SWOT

The Pythons	MATRIZ SWOT – FOFA	
	Fatores Positivos	**Fatores Negativos**
Fatores Internos	**Forças** -Produto rápido e eficiente; - Método para atrair pessoas jovens, suprindo a falta das mesmas; - Suprir uma demanda da TV Gazeta, dado a falta de uma ferramenta parecida.	**Fraquezas** - Possíveis margens de erro; - Falta de uma predição precisa, caso haja programações atípicas.
Fatores Externos	**Oportunidades** - Pode expandir o alcance da rede Gazeta; - Poder diminuir a demanda manual das predições; - Criar programas mais assertivos de acordo com o publico alvo e horário; - Aumentar o tempo de permanencia dos espectadores	**Ameaças** - Margem de erro dado espontaneidade humana; -Falta de demanda da TV; -Mar vermelho, grande concorrência de mercado.

4.1.3. Planejamento Geral da Solução

a) Dados Disponíveis:

1. **Mapeamento de Audiência:** Planilha do Excel contendo informações sobre a audiência para o canal da Rede Gazeta e outros canais para comparação, assim como a audiência geral dos Canais Pagos e Serviços Não Identificados *(Streamings)*.

2. **Grade Horária de Programação:** Grade contendo quais programas estão sendo transmitidos em relação a horário e mês, assim como as categorias em que esses programas se encaixam.

b) Qual a solução proposta

A solução se trata de um modelo preditivo, onde o usuário possa inserir no modelo um possível horário para um novo programa de TV. Assim, o modelo pode fornecer dados sobre a audiência daquele programa, como gênero dos telespectadores, faixa etária e tipo de programa que esse público tem preferência por assistir.

Por fim, auxiliará os produtores a criarem um programa que se adapte mais à possível audiência naquele horário e os profissionais da publicidade a organizarem campanhas de marketing mais assertivas, além de possibilitar a aplicação de investimentos com maior consciência dos resultados.

c) Qual o tipo de tarefa (regressão ou classificação):

O tipo de tarefa da predição a ser desenvolvido, inicialmente, é regressão, visto que trabalharemos com dados contínuos e receberemos um feedback de probabilidade. No entanto, a classificação também seria um caminho possível e interessante caso houvessem parâmetros de sucesso de audiência (baixo, médio e alto, por exemplo) a partir dos dados gerados pela predição.

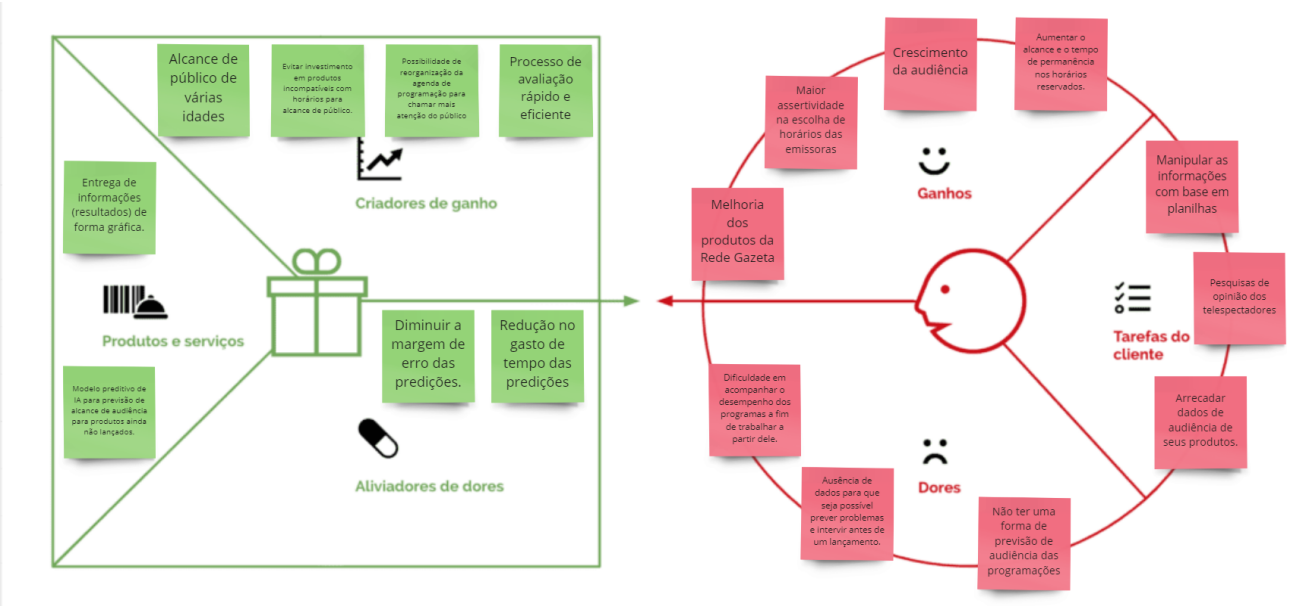
d) Como a solução proposta deverá ser utilizada:

O modelo preditivo desenvolvido tem o objetivo de realizar uma previsão com uma boa taxa de precisão sobre o quanto uma programação pode obter sucesso ou a baixa, subindo ou caindo a taxa de audiência e o tempo em que o telespectador fique no mesmo canal.

e) Nosso produto visa impulsionar as estatísticas televisivas da TV Gazeta, em duas frentes, sendo a primeira em pico de telespectadores totais, seja em filmes, novelas, noticiários, etc. Já a segunda frente o tempo de retenção e permanência do telespectador durante a exibição do programa.

f) Critério de Sucesso: Atualmente podemos definir os critérios de sucesso do negócio tendo como base objetivo específico do projeto fornecido pelo cliente, no qual devemos prever a audiência em um período específico de tempo e com base nesse resultado pode prever, quanto investimento ele terá que colocar para o programa atingir a expectativa de telespectadores, aumentar a audiência e o tempo de permanência. Consideramos que obtivemos sucesso ao atingir o objetivo, e isso pode ser metrificado por meio da existência de uma margem de erro pequena em relação a audiência prevista em comparação com o valor real. Quando falamos de margem de erro pequena se torna necessário definir inicialmente o que é uma margem pequena ou não em relação ao modelo de negócios. Para isso cabe entendermos com o parceiro de negócios qual o nível de precisão desejado.

4.1.4. Value Proposition Canvas



Proposta de valor:

Produtos e serviços:

- Entrega de informações (resultados) de forma gráfica.
- Modelo preditivo de IA para previsão de alcance de audiência para produtos ainda não lançados.

Criadores de ganho:

- Alcance de público de várias idades
- Evitar investimento em produtos incompatíveis com horários para alcance de público.
- Possibilidade de reorganização da agenda de programação para chamar mais atenção do público
- Processo de avaliação rápido e eficiente

Aliviadores de dores:

- Diminuir a margem de erro das predições.
- Redução no gasto de tempo das predições

Perfil do cliente:

Ganhos:

- Crescimento da audiência.
- Aumentar o alcance e o tempo de permanência nos horários reservados.
- Maior assertividade na escolha de horários das emissoras.
- Melhores investimentos dos produtos da Rede Gazeta.

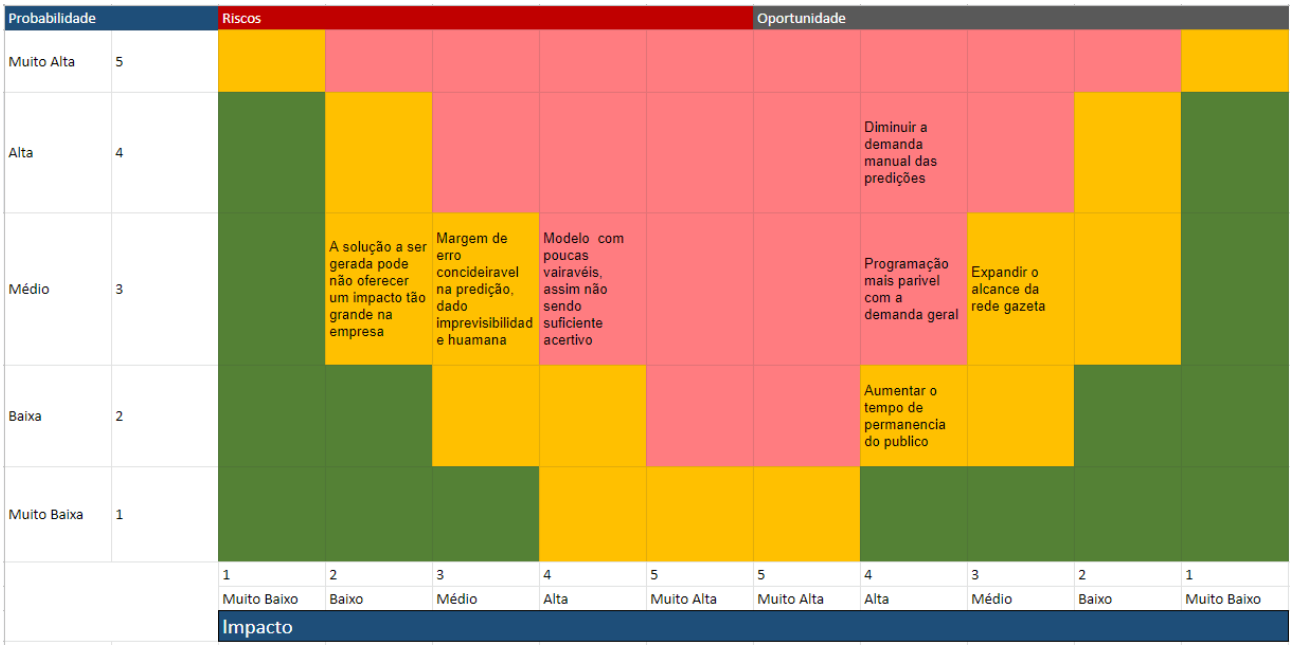
Tarefas do cliente:

- Manipular as informações com base em planilhas
- Pesquisas de opinião dos telespectadores
- Arrecadar dados de audiência de seus produtos.

Dores:

- Dificuldade em acompanhar o desempenho dos programas a fim de trabalhar a partir dele.
- Ausência de dados para que seja possível prever problemas e intervir antes de um lançamento.
- Não ter uma forma de previsão de audiência das programações

4.1.5. Matriz de Riscos



4.1.6. Personas

Nome
Cassia Ellen

Cargo
Analista de Dados

Era
25 to 34 years

Nível mais alto de educação
Mestrado em Tecnologia

Indústria
Tecnologia

Tamanho da organização
+500 Funcionários

Características

- **Gostos:** Rock Clássico, Heavy Metal, Pagode e filmes do Tarantino.
- **Personalidade:** Cassia é muito obstinada e gosta de otimizar seu tempo e processos no trabalho para que as ações sejam realizados de forma automatizada.
- **Hobbies:** Ler, aprender sobre Dados, ver filmes, passar tempo com sua familia.
- **Vida Pessoal:** Cassia tem 2 filhos e gostaria muito de dedicar mais tempo com sua familia, por conta disso tem interesse em otimizar seu trabalho.

Responsabilidades do trabalho

- Analisar os dados de audiência fornecidos pelo pela Kantar Ibope.
- Fornecer relatórios sobre o alcance de audiência em relação aos parâmetros fornecidos pela Kantar Ibobe.

Ferramentas que eles precisam para fazer seu trabalho

- Situação Atual: Excel contendo a base dos dados.
- Situação Futura: Agamotto - Modelo de previsão de dados indicando o score de audiência.

Maiores desafios

- Obter uma análise de dados em um grande volume com base em um histórico de evidências.
- Automatização e Praticidade no processo de gerar dados de referência
- Conseguir ter uma previsão antecipada de quanto de audiência o programa possivelmente irá ter.

4.1.7. Jornadas do Usuário

Cassia Ellen

Cenário e Expectativa : Gerar melhores opções estratégicas com base em uma predição de alcance de público de um produto em determinado horário.

FASE 1 **Recebimento dos Dados**	FASE 2 **Alimentação da Predição**	FASE 3 **Aplicação do Modelo**	FASE 4 **Análise de Dados Obtidos**
1. **Receber dados de audiência fornecidos anualmente pela IBOPE;** 2. **Mapear grade horária respectiva da emissora para cada horário;** 3. **Mapear a categoria do programa.** 4. **Reunir e organizar os dados em um banco centralizado para serem utilizados na predição, inserindo cada uma das planilhas disponibilizadas (Grade Horária, Audiência por Hora).**	1. **Inserir os bancos de dados do histórico de audiências e programas (Grade Horária) para obter a predição desejada.**	1. **Obter os resultados a partir do modelo de predição definido e assim conseguir adquirir o score de audiência**	1. **Analisar resultados: Entender como os dados fornecidos pelo modelo podem ser aplicados nos setores de estratégia para a emissora.** 2. **Identificar a taxa de audiência para o horário e programa analisado: Entender como cada um dos fatores impacta determinado programa e como podem ser otimizados para a evolução do score - possibilidade de sucesso para o programa.**
Preciso conseguir entender os dados fornecidos pelo Ibope e entender a relação entre eles	Preciso conseguir fornecer os bancos necessários para os quais quero a analise de score	Quero poder obter o score de audiência e ter a informação de quais parâmetros são mais relevantes dentro de cada horário	Quero obter dados suficientes para criar propostas e hipóteses de orientações estratégicas para que meus superiores possam gerar melhorias no negócio.

Oportunidades

- Otimizar o input de Dados para inserção do Banco de Dados

Responsabilidades

- Criar um processo otimizado de inserção de banco de dados.
- Garantir que estamos fornecendo os resultados pertinentes para a definição do score de audiência.
- Mostrar o quanto cada feature influencia no resultado final da audiência.

4.2. Compreensão dos Dados

1. *Descrição dos dados a serem utilizados:*

    a. Descrição dos Dados:

        i. Fonte dos Dados: Kantar Ibope, 2022

        ii. Tipo de Arquivo: Planilha - CSV/ XLSX

        iii. Tamanho da Base de Dados: 427MB

        iv. Quantidade de Planilhas: 18 Planilhas
            - Canais Pagos (3 Planilhas [Sab., Dom., Seg - Sex])
            - NI Conteúdo (3 Planilhas [Sab., Dom., Seg - Sex])
            - TLE (3 Planilhas [Sab., Dom., Seg - Sex])
            - TV 0 (3 Planilhas [Sab., Dom., Seg - Sex])
            - TV 1 (3 Planilhas [Sab., Dom., Seg - Sex])
            - TV 2 (3 Planilhas [Sab., Dom., Seg - Sex])

        v. Colunas: 60 Colunas: Cada coluna representa as seguintes características para cada parâmetro de avaliação. Sendo que os parâmetros são dados por

            - Rat% (Rating): Número de indivíduos por domicílio acompanhando determinado evento → Refere se a audiência.
            - Shr% (Share): Participação da audiência em um evento, em relação ao total de aparelhos ligados.
            - Rch% (Reach): Total de indivíduos diferentes atingidos por pelo menos 1 minuto por um conjunto de eventos.
            - Fid% (Fidelidade):

            Características avaliadas para cada parâmetro:

            - Gênero: Feminino e Masculino;
            - Classe Social: AB, C1, C2, DE;
            - Faixa Etária (anos): 4-11, 12-17, 18-24, 25-34, 35-49, 50-59, 60+.

        vi. Tipos de Dados: Datetime (dia e horário), String (texto), Inteiro (número).

        vii. Período de Análise: 2020 - 2022

b. Mesclando os dados, fizemos a confecção de três tabelas, a primeira medindo a relação entre audiência e idade, identificando quais faixas de idade apresentam maiores picos de audiência. A segunda identificando a relação de audiência e gênero, podendo visualizar a porcentagem de telespectadores dividida entre público masculino e público feminino. Por último, uma tabela geral, que pega a média de audiência dos dias da primeira semana apresentada no banco de dados, e depois categorizando essa média em idade e gênero.

c. O risco que temos ao lidar com esses dados é o fato de que os dados da audiência são provenientes do IBOPE e a mesma base não pode ser compartilhada e deve ser anonimizada. Além disso, a diversidade não é tão grande, visto que temos acesso a grade de programação de apenas duas outras emissoras, gerando uma pequena base de dados para comparar a relevância do tipo de programa com a sua audiência.

d. Utilizando dos subconjuntos para visualizar o nível de audiência, divididos tanto pela faixa de idade, quanto pelo gênero, nos provê uma indicação consistente de quais tipos de programas atraem quais tipos de pessoas e utilizaremos isso como base para medir índice de audiência de um título que ainda não foi passado, ou qual seria o êxito de um título repetido.

e. Os dados disponibilizados para estudo apresentam restrição de segurança para preservar a privacidade dos dados das emissoras que foram usadas para fins de comparação. Sendo assim, não podem ser disponibilizados publicamente.

2. Descrição Estatística Básica dos Dados:

Números dos dados: Baseado na base de dados , classificamos as estatísticas da seguinte forma:

Média:
Faixa etária:

- 4-11: 2.23;
- 12-17: 2.51;
- 18-24: 3.63;
- 25-34: 3.77;
- 35-49: 4.41;
- 50-59: 6.19;
- 60+: 8.10

Mediana:
Faixa etária:

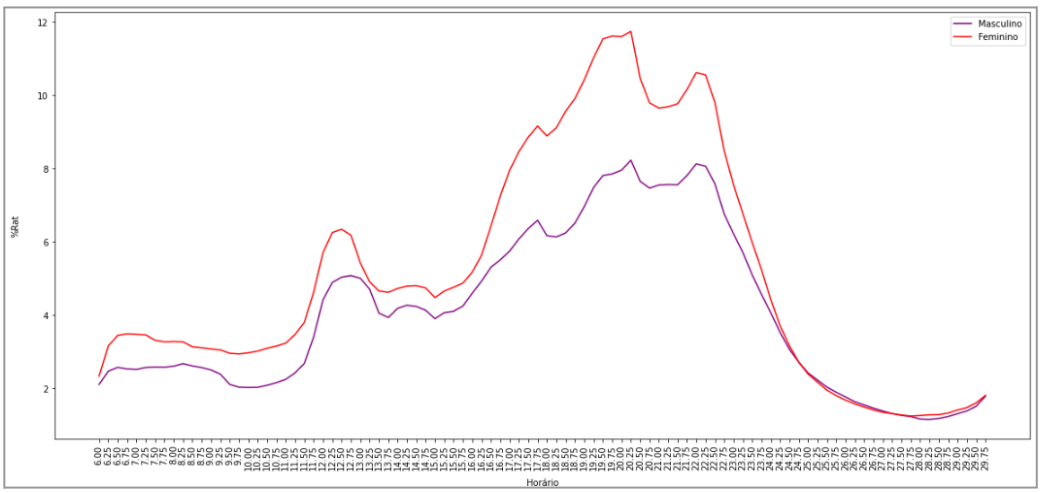
- 4-11: 1.47;
- 12-17: 1.57;
- 18-24: 2.61;
- 25-34: 2.89;
- 35-49: 3.78;

- 50-59: 5.01;
- 60+: 6.68

Desvio padrão:
Faixa etária:
- 4-11: 2.49967781;
- 12-17: 2.909038769;
- 18-24: 3.639758362;
- 25-34: 3.349938817;
- 35-49: 3.095652657;
- 50-59: 4.892826927
- 60+: 6.032953907

Atributos de interesse (Tipos de dados):

a. Análise: Gênero X Horário
   i. Média:
      - Masculino: 4.03
      - Feminino: 5.14;
   ii. Mediana:
      - *Masculino:* 3.41
      - Feminino: 4.08;
   iii. Desvio Padrão:
      - Masculino: 2.785044321
      - Feminino: 3.808778465;



Benefício da Análise: Entender qual o gênero de maior prevalência de forma macro em relação ao impacto na audiência e também poder visualizar como os gêneros podem ser um critério de padrão de audiência para cada horário.

b. Análise: Classe Social X Horário
   i. *Média:*
      - *AB:* 4.03;
      - C1: 3.66;
      - C2: 5.17

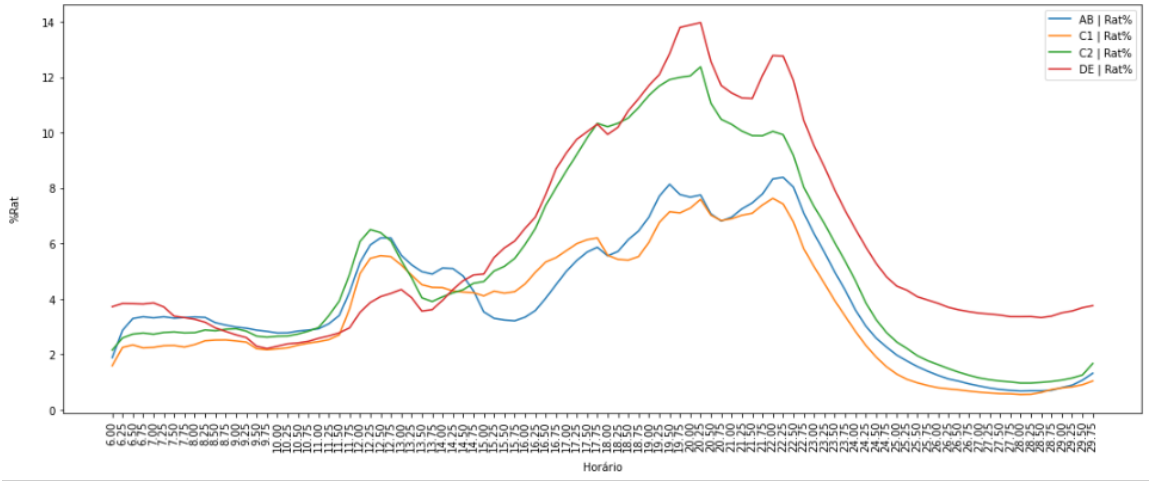
- DE: 6.05;

ii. Desvio Padrão:

- AB: 2.903602071;
- C1: 3.151994451;
- C2: 4.207737178;
- DE: 4.985268943;

iii. Mediana:

- AB: 3.51;
- C1: 2.96;
- C2: 3.95;
- DE: 4.7;



Benefício: Entender qual a classe social de maior prevalência de forma macro em relação ao impacto na audiência e também poder visualizar se há uma variação entre a preferência de cada classe para determinarmos um critério de padrão de audiência para cada horário.

c. Análise: Faixa Etária X Horário:

i. *Média:*

- 4-11: 2.23;
- 12-17: 2.51;
- 18-24: 3.63;
- 25-34: 3.77;
- 35-49: 4.41;
- 50-59: 6.19 e
- 60+: 8.10

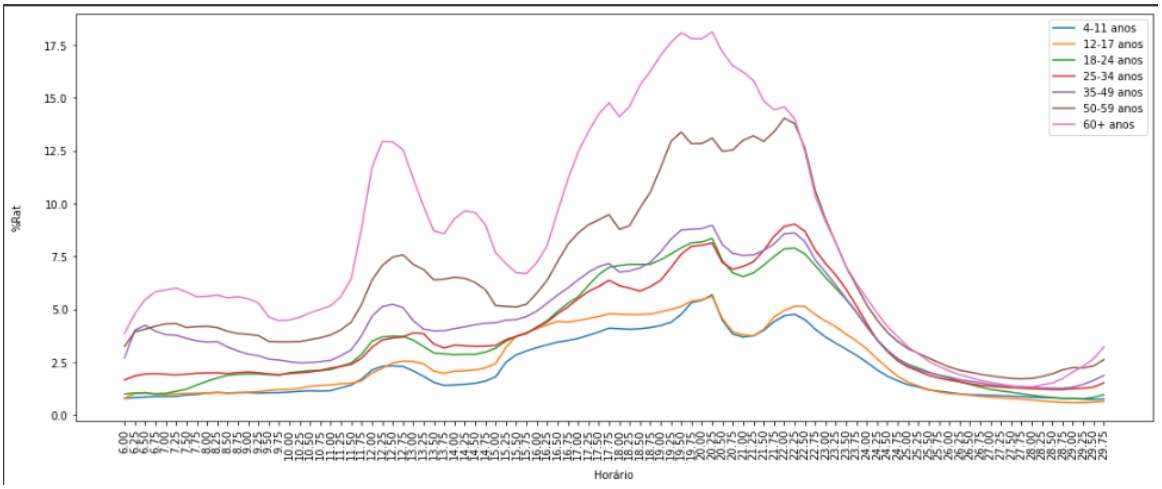
ii. Mediana

- 4-11: 1.47;
- 12-17: 1.57;
- 18-24: 2.61;

- 25-34: 2.89;
- 35-49: 3.78;
- 50-59: 5.01
- 60+: 6.68

iii. Desvio Padrão:

- 4-11: 2.49967781;
- 12-17: 2.909038769;
- 18-24: 3.639758362;
- 25-34: 3.349938817;
- 35-49: 3.095652657;
- 50-59: 4.892826927
- 60+: 6.032953907

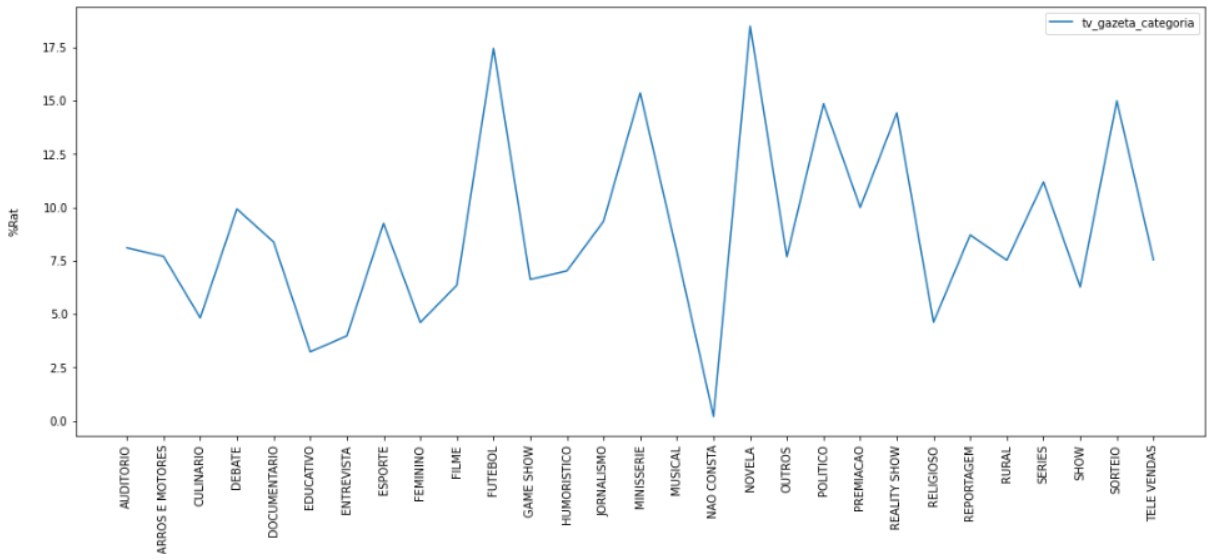


Benefício: Entender qual a faixa etária de maior prevalência de forma macro em relação ao impacto na audiência e também poder visualizar se há uma variação entre a preferência de cada idade para determinarmos um critério de padrão de audiência para cada horário.

d. Análise: Rede Gazeta em Relação a Concorrentes

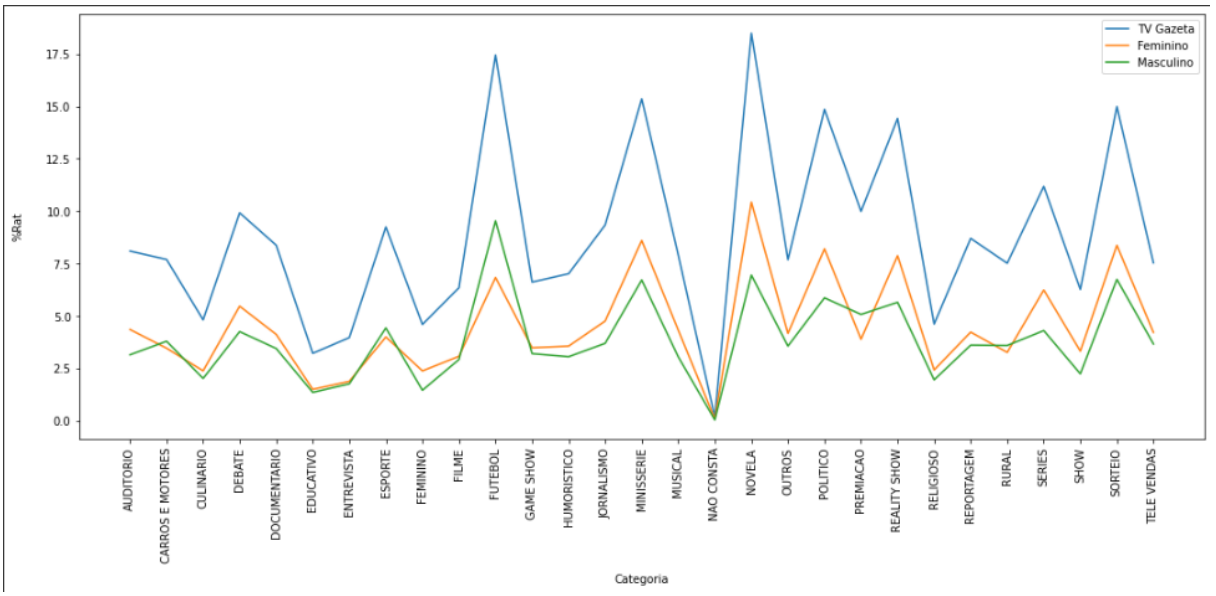
Benefício: Comparação entre os horários de maior audiência podendo visualizar os dados de audiência em relação a Rede Gazeta e a outros canais para entendermos se os momentos em que temos uma menor audiência na TV Gazeta implica em uma maior audiência em outros canais ou são outros critérios que norteiam a redução da audiência.

e. Análise: Audiência X Categoria



Benefício: Visualização de qual categoria de programa possui uma maior audiência com base nos dados dos últimos 3 anos. Assim podemos ter uma boa métrica de quais programas ou serviços mais alcançam o público.

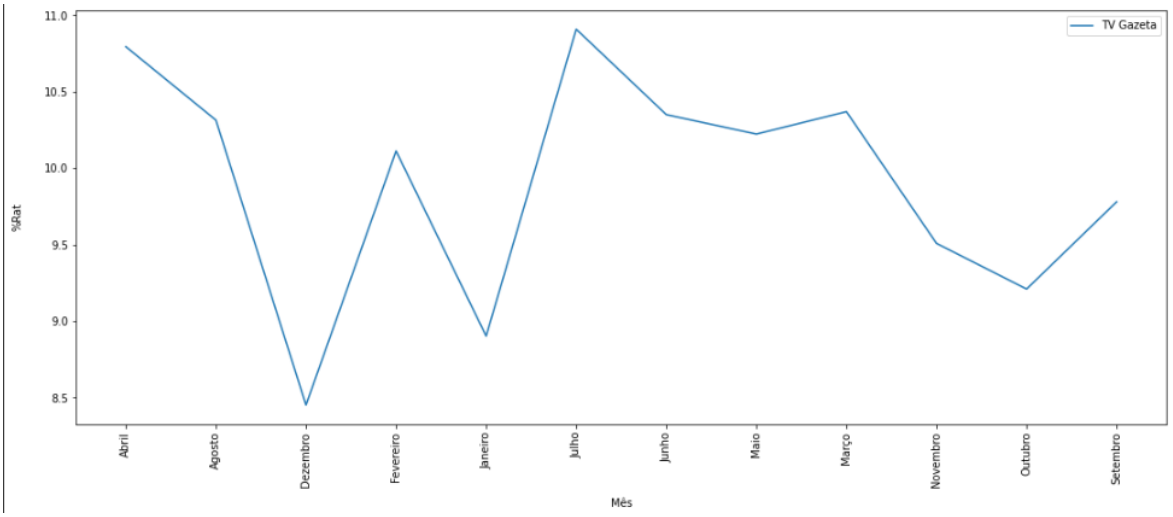
f. Análise: Gênero X Categoria



Beneficio: Relação entre a categoria do programa transmitido e gêneros predominantes em cada uma das categorias. Nos permite validar novamente a

questão de que é um critério de impacto na audiência e também ter algumas hipóteses sobre quais tipos de programas devem considerar mais ou menos o gênero que deverá ser o público alvo para as chamadas de marketing, por exemplo. Para os programas em que o gênero não varia muito em relação a categoria, podemos entender que esse critério é de menor peso para esse tipo de programa. Pois o público é unificado em relação ao gênero.

g. Análise: Audiência (Total de Domicílios) X Mês



Benefício: Visualização dos meses nos quais existem maiores ou menores picos de audiência. O que posteriormente pode ser conectado com o tipo de programa (Categoria) ou até mesmo o evento que está sendo transmitido nesse período, assim garantimos que nossa visualização terá parâmetros generalizados e não apenas eventos específicos que impactam a audiência pontualmente.

Atributos de interesse (Visualização):

Para todas as situações analisamos a relação entre o parâmetro pré-determinado no item acima por meio do Rat% de Audiência por Domicílio. Assim tivemos sempre um mesmo parâmetro de referência para poder criar uma intersecção entre dados de um mesmo parâmetro.

3. *Descrição da predição desejada ("target"), identificando sua natureza (binária, contínua, etc.)*

O target da predição é a coluna "audiência", já que o software dará como resultado uma predição de score de audiência, taxa de permanência e alcance de público de um programa a ser lançado em determinado horário, definidos a partir dos dados de audiência previamente coletados.

O modelo consiste na entrada de um conjunto de exemplos (dados) que, como mencionado anteriormente, será a audiência com base no histórico e que se baseia em rótulos de valores

conhecidos. E temos como objetivo uma saída de um algoritmo de regressão que é uma função, que será usado para prever o valor de rótulo para qualquer novo conjunto de recursos de entrada.

Os dados da coluna de rótulo de entrada devem ser sempre do tipo Float. No nosso caso aplicamos a audiência como rótulo principal de entrada buscando definir como as características dos telespectadores.

Os treinadores para esta tarefa produzem a saída contendo a predição desejada de audiência e o quanto cada uma das variáveis (coeficientes de uma função, por exemplo) podem impactar no valor final de audiência.

4.3. Preparação dos Dados

4.3.1 Mesclagem das tabelas:

Para conseguirmos unificar as informações em uma mesma tabela na qual centralizamos as informações necessárias para rodarmos os modelos de regressão, realizamos um processo de mesclagem:

Link para o código de mesclagem.

Passo 1: Definição dos Dados

Selecionamos a tabela "Histórico.csv" fornecida pelo parceiro de negócios contendo as informações disponibilizadas pelo Kantar Ibope referentes ao histórico. Em seguida, selecionamos a tabela "Grade Horária.csv" que contém a grade horária de todas as emissoras que estamos trabalhando em nossa modelagem de dados assim como as categorias determinadas pelo Kantar Ibope.

Passo 2: Mesclagem dos Dados (Colab)

Para realizarmos o processo de união em relação a esses dados inicialmente definimos os intervalos dos bancos nos quais iremos inserir a informação de qual programa está passando naquele período de tempo e qual é a categoria dessa programa de acordo com o Kantar Ibope.

Em seguida selecionamos quais bibliotecas seriam necessárias para realizarmos a mesclagem:

```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from datetime import datetime
```

Feita a seleção e importação das bibliotecas selecionadas para a aplicação da mesclagem, foram importadas as tabelas referentes a:

- Grade Horária (Programas da: TV 0, TV 1 e TV 2)
- Histórico de Audiência (Dias da Semana, Sábado e Domingo para TV 0, TV 1 e TV 2)

Temos então uma função que recebe um parâmetro genérico como o nome de uma emissora e nos fornece o banco de dados que mescla as informações de horários em relação a cada um dos programas e categorias.

Em seguida foi feita uma concatenação entre os dias da semana, sábado e domingo. Isso aplicado para o parâmetro "emissora" que é uma entrada da função.

O próximo passo foi criar um novo banco no qual iremos ter os dados de todos os dias concatenados e passamos as colunas de "Data" para o formato "*Datetime*" da Biblioteca Pandas utilizada.

Feitas as concatenações e a passagem para o formato *"Datetime"* podemos ordenar os dados de acordo com a data. Feito isso, conseguimos realizar a mesclagem entre a planilha de grade horária e a de histórico de programas tendo como base de mesclagem a data e hora do programa.

Por fim, a função nos gera um .csv contendo uma versão final do banco com os dados todos mesclados em seus respectivos dias, horários e as audiências correspondentes a cada programa.

4.3.2. Anonimização dos dados:

Alteramos os nomes das emissoras para números ordenados (tv_0, tv_1, tv_2), a fim de manter a confidencialidade dos dados e pela necessidade de descrição ao usar dados externos para comparação. Por motivos de anonimidade, também convertemos todos os programas para valores quantitativos, os relacionando com as suas emissoras.

Exemplo: o programa 1, da TV 0, foi convertido em *"programa 0.1"*. Já para o Programa 1 da TV 1, foi convertido para "*programa 1.1"*, e para o Programa 1 da TV 2, convertemos em "*programa 2.1"*. O mesmo padrão foi seguido para os outros programas de cada uma das emissoras.

4.3.3 Feature Engineering:

Padronização dos dados:

Foi necessário converter as colunas "Mês" e "Dia da Semana", que originalmente são strings, para valores quantitativos, possibilitando a manipulação.

Dessa forma, os dados foram atribuídos da seguinte maneira: Janeiro : 1, Fevereiro: 2, Março: 3, Abril: 4, Maio: 5, Junho: 6, Julho: 7, Agosto: 8, Setembro: 9, Outubro: 10,

Novembro: 11, Dezembro: 12. Em relação aos dias da semana, também seguimos um padrão de numeração: Segunda: 1, Terça: 2, Quarta: 3, Quinta: 4, Sexta: 5, Sábado: 6, Domingo: 7.

Em relação ao horário, fizemos a representação para análise por quarto de hora, ou seja, a cada 15 minutos. Assim, temos uma organização horária indicando, por exemplo: 06:00:00, 06:05:00 e 06:10:00 foram convertidos para 6, 06:15:00, 06:20:00 e 06:25:00 foram convertidos para 6.25 (usando os horários em modelo hh:mm:ss e os números de identificação convertidos como quarto de cem, pois a conversão é em formato de número e não de hora) até o último horário da base de dados, que se encerra com a conversão de 29.75, equivalente ao horário 29:55:00.

Valores ausentes ou em branco:

Não foi necessária a realização de nenhuma manipulação dos dados no intuito de tratar os valores ausentes ou nulos, visto que não há dados faltantes.

Seleção dos dados (colunas):

A planilha original foi manipulada de forma a manter somente as colunas que serão utilizadas para criarmos as diferentes regressões e hipóteses. Como decidimos trabalhar, inicialmente, com o Rat, retiramos as outras colunas que não tinham relação com o mesmo. Sendo assim, os dados mantidos foram: Data, Hora Início, Emissora, Mês, Dia do Mês, Dia da Semana, Total Domicílios, colunas de Rat para cada classe social, coluna de Rat para cada faixa de idade, Programa, Categoria e Faixa

Além disso, foi criada a coluna Horário, está é feita com base na divisão dos horários em quartos de hora de acordo com a padronização dos dados que já havia sido estabelecida antes e segue os seguintes parâmetros:

- Manhã 1 : 6.0 - 9.75
- Manhã 2: 10.0 - 11.75
- Almoço: 12.0 - 14.75
- Tarde: 15.0 - 17.75
- Noite 1 : 18.0 - 20.75
- Noite 2 : 21.0 - 24.5
- Madrugada: 24.75 -29.75

Relação de colinearidade

Neste caso escolhemos as 3 emissoras e rodamos um modelo que calcula a correlação entre todas as variáveis, a partir disso podemos ter alguns possíveis parametros para rodar a regressão linear.

https://colab.research.google.com/drive/1gmglZnEj52AHufEtbdusmEGrZcB0Y4KZ?usp=sharing

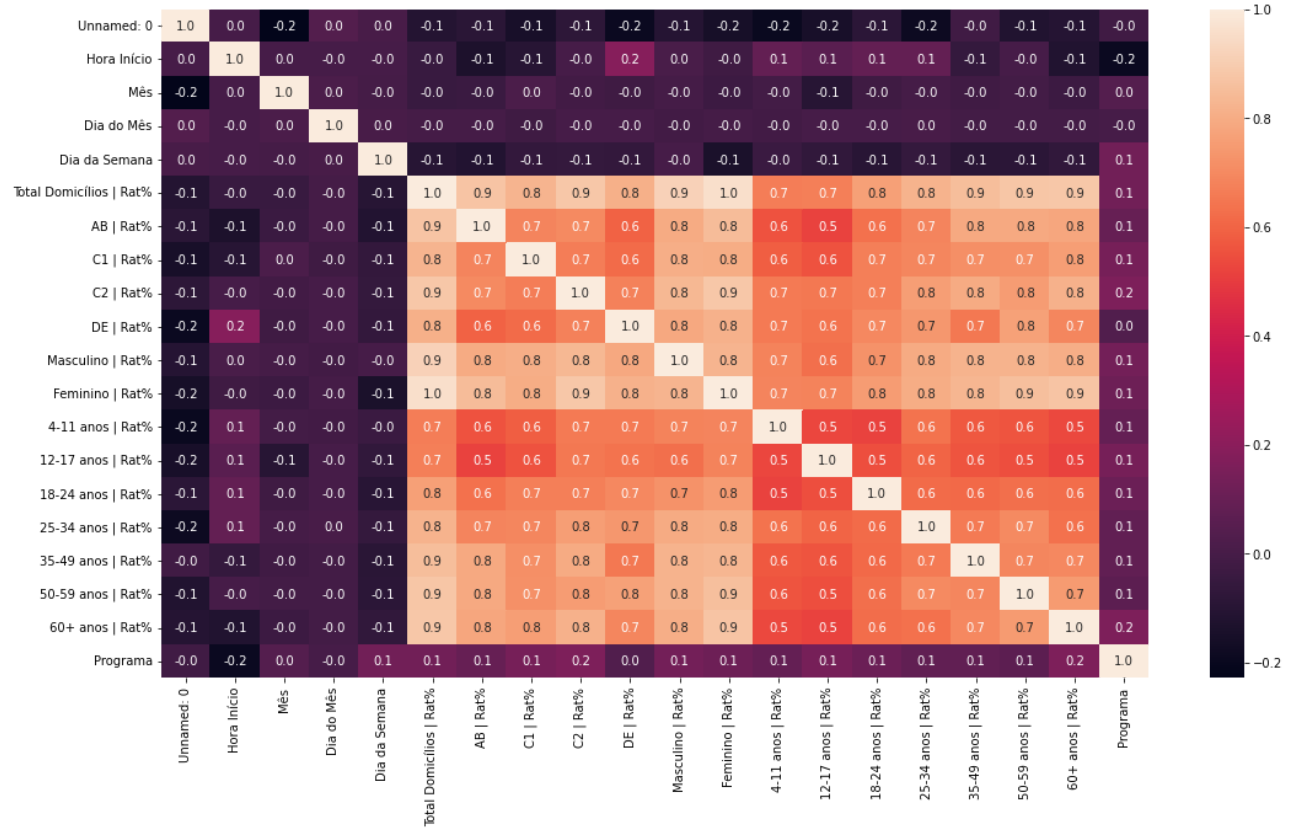


Figura que representa a TV_0

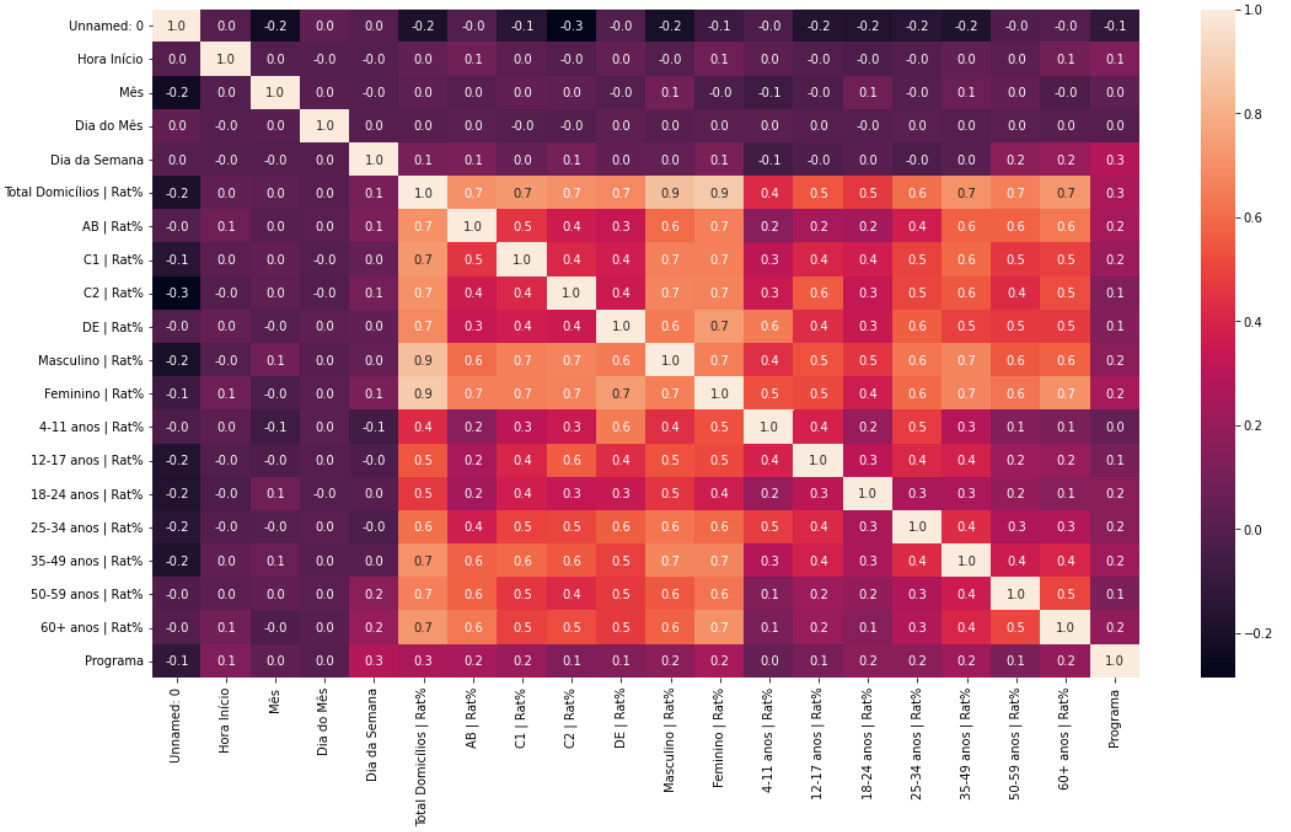


Figura que representa a TV_1

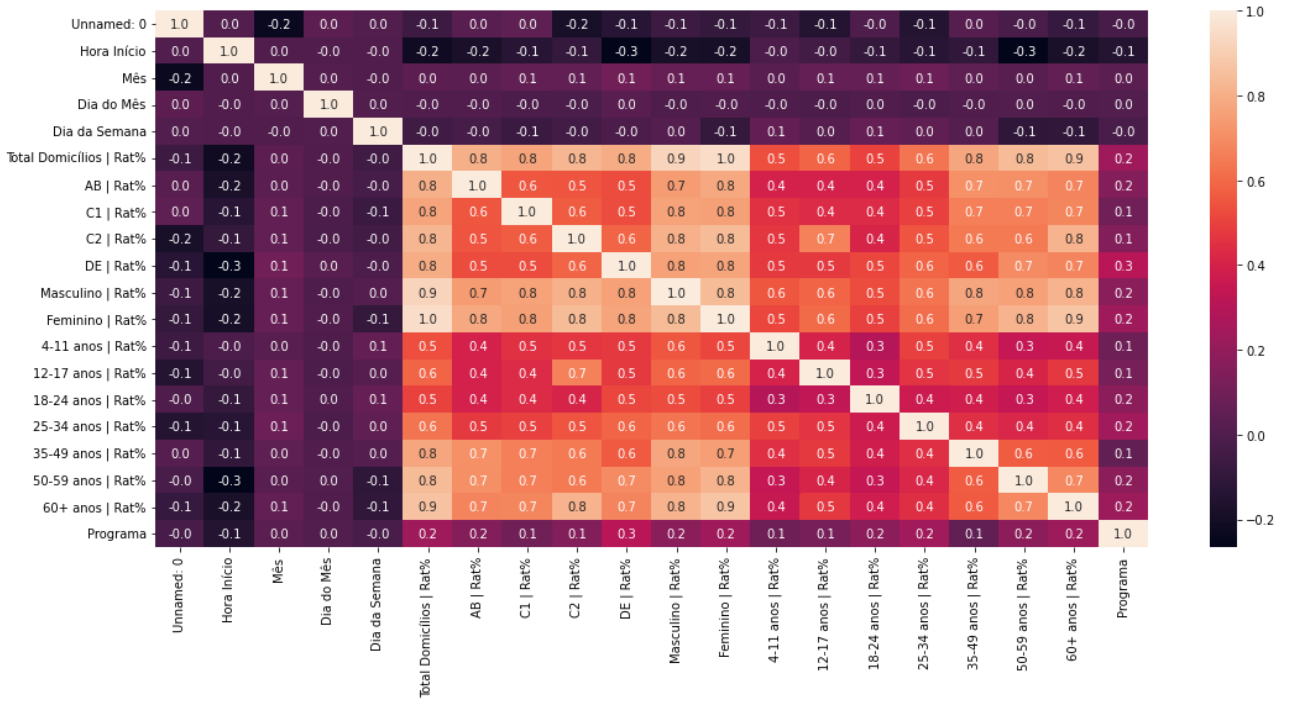


Figura que representa a TV_2

De acordo com o gráfico quanto mais próximo de 1.0 maior a correlação entre as variáveis representadas no gráfico.

4.4. Modelagem

Para a Sprint 3, você deve descrever aqui os experimentos realizados com os modelos (treinamentos e testes) até o momento. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus experimentos e resultados.

Para a Sprint 4, você deve realizar a descrição final dos experimentos realizados (treinamentos e testes), comparando modelos. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus experimentos e resultados.

4.5. Avaliação

Nesta seção, descreva a solução final de modelo preditivo, e justifique a escolha. Alinhe sua justificativa com a seção 4.1, resgatando o entendimento do negócio e explicando de que formas seu modelo atende os requisitos. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus argumentos.

4.6 Comparação de Modelos

5. Conclusões e Recomendações

Escreva, de forma resumida, sobre os principais resultados do seu projeto e faça recomendações formais ao seu parceiro de negócios em relação ao uso desse modelo. Você pode aproveitar este espaço para comentar sobre possíveis materiais extras, como um manual de usuário mais detalhado na seção “Anexos”.

Não se esqueça também das pessoas que serão potencialmente afetadas pelas decisões do modelo preditivo, e elabore recomendações que ajudem seu parceiro a tratá-las de maneira estratégica e ética.

6. Referências

Nesta seção você deve incluir as principais referências de seu projeto, para que seu parceiro possa consultar caso ele se interessar em aprofundar.

Utilize a norma ABNT NBR 6023 para regras específicas de referências. Um exemplo de referência de livro:

LUCK, Heloisa. **Liderança em gestão escolar**. 4. ed. Petrópolis: Vozes, 2010.

SOBRENOME, Nome. **Título do livro**: subtítulo do livro. Edição. Cidade de publicação: Nome da editora, Ano de publicação.

Anexos

Utilize esta seção para anexar materiais como manuais de usuário, documentos complementares que ficaram grandes e não couberam no corpo do texto etc.