O Oráculo
TV Gazeta

inteli
instituto
de tecnologia
e liderança

Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
11/08/2022	Yasmin	1.0	1. Introdução
11/08/2022	Gabrio e Patricia	1.0	4.1.2. Matriz SWOT e 4.1.5. Matriz de Riscos
11/08/2022	Gustavo	1.0	4.1.1. Contexto da indústria
11/08/2022	Gabrio, Gustavo, Patricia, Tainara, Yasmin e Victor	1.0	4.1.3. Planejamento Geral da Solução
11/08/2022	Gustavo e Tainara	1.0	4.1.4. Value Proposition Canvas
11/08/2022	Gabrio, Gustavo, Patricia, Tainara, Yasmin e Victor	1.0	4.2. Compreensão dos Dados
12/08/2022	Yasmin	1.1	6. Referências
15/08/2022	Gabrio, Gustavo, Yasmin, Tainara, Victor, Patricia	2.0	4.1.6. Personas
17/08/2022	Gabrio, Yasmin, Patricia	2.0	4.1.7 Jornada do Usuário
26/08/2022	Tainara e Victor	2.1	4.3.2 Manipulação de dados e registros
26/08/2022	Gustavo e Patricia	2.1	4.3.2 Agregação de Registros e derivação de novos atributos
26/08/2022	Gustavo, Tainara e Yasmin	2.1	4.3.4 Remoção e substituição de valores ausentes, em branco, ou desconsiderados
26/08/2022	Gustavo, Tainara e Yasmin	2.1	4.3.5 Identificação das features selecionadas
27/08/2022	Gábrio, Patricia	2.1	4.3.1 Anonimização dos dados
28/08/2022	Gábrio	2.2	Revisão, padronização e formatação do documento.

Sumário

1. Introdução

Apresente de forma sucinta o parceiro de negócio, seu porte, local, área de atuação e posicionamento no mercado. Maiores detalhes deverão ser descritos na seção 4

Descreva resumidamente o problema a ser resolvido (sem ainda mencionar a solução).

Caso utilize citações ao longo desse documento, consulte a norma ABNT NBR 10520. Sugerimos o uso do sistema autor-data para citações.

A medição de audiência na televisão acontece por meio da tradicional amostragem com o aparelho chamado People Meter em um lar escolhido, este possui mapas de calor para identificar se a pessoa está acompanhando determinada mídia e quem participa da pesquisa tem seu próprio número, informando o sexo, idade, classe economica e programação assistida. Embora o PeopleMeter tenha o principal objetivo de medir o índice de audiência, verificando o tamanho e a composição do público que acompanha determinada programação, por meio desse método, as informações são insuficientes para a medição do sucesso dos programas de TV, tal como sua classificação antecipada e se o público atenderá as expectativas da audiência da programação, ou seja, em razão das limitações da tecnologia existente, há dificuldade no entendimento e previsão do comportamento da audiência, tendo grandes chances de gerar prejuízos financeiros, mobilização frequente de profissionais e equipes devido a atualizações e cancelamentos da programação, qualidade da programação e, claro, a perda de espectadores.

Desse modo, não há uma visão 360° para isolar os vários fatores que levam ao declínio de audiência, dado que, todos os dias os espectadores são inundados com diversas opções de entretenimento, havendo uma necessidade de informações do que está funcionando e a razão que faz o espectador não assistir determinada programação ou mudar de canal, preferindo a concorrência. A exibição da programação com conteúdos selecionados de acordo com o que interessa quem mais importa, o público, são escalados sem uma base preditiva consistente nas decisões sobre produção e cronogramas de programas, alterando a classificação e visão do canal, principalmente ao se tratar da atualidade, em que há intensa competição entre canais devido o surgimento de novas mídias.

2. Objetivos e Justificativa

2.1. Objetivos

Descreva resumidamente os objetivos gerais e específicos do seu parceiro de negócios

2.2. Justificativa

Faça uma breve defesa de sua proposta de solução, escreva sobre seus potenciais, seus benefícios e como ela se diferencia.

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

Descreva brevemente a metodologia CRISP-DM e suas etapas de processo

3.2. Ferramentas

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

3.3. Principais técnicas empregadas

Descreva brevemente as principais técnicas empregadas, algoritmos e seus benefícios

4. Desenvolvimento e Resultados

4.1. Compreensão do Problema

4.1.1. Contexto da indústria

A TV aberta possui um mercado que conta com diversos players com atuação intensa, entre os principais podemos citar: Rede Globo, o maior conglomerado de mídia e comunicação da américa latina e a segunda maior do mundo, a Record, considerada a segunda maior emissora do Brasil e a quinta maior do mundo, o SBT, uma emissora fundada pelo empresário e animador Silvio Santos e por fim a Bandeirantes, fundada pelo empresário João Jorge Saad e sendo considerada a quarta maior rede de televisão do país.

As emissoras abertas oferecem para o público uma variedade de programas de forma gratuita. O foco das emissoras que possuem esse modelo de negócio é capturar a atenção dos telespectadores para vendê-la aos anunciantes e assim gerar rentabilidade. Consequentemente, os produtos desenvolvidos tendem a ser genéricos buscando alcançar uma alta taxa de aceitabilidade entre o maior número de telespectadores.

A atual infraestrutura tecnológica provoca um gargalo para a inovação da TV aberta. Uma das primeiras estratégias utilizadas foi a preservação da qualidade de imagem superior às outras mídias. Entretanto, com o aumento da qualidade em dispositivos, redes sociais e serviços de streaming concorrentes, a rede televisiva se viu frente a um problema em que se seria necessário buscar novas estratégias. A convergência de diversos serviços para uma única plataforma digital é um movimento atual das grandes emissoras e teve um efeito de encadeamento, sendo adotado pelos grandes players. O público mais jovem consome conteúdos mais curtos e sucintos, sendo essa a tendência presente no mercado midiático. Portanto, a televisão deve expandir o seu alcance para uma diversidade de aparelhos maior e levar em consideração o contexto do consumo, colocando em cartaz conteúdos com diferentes tipos de duração.

Para abordarmos uma melhor percepção do contexto do mercado, analisaremos o cenário de atuação da empresa parceira. A ferramenta escolhida foram as 5 forças de porter.

Poder de negociação dos fornecedores: Fornecedores são quem provêm insumos para criação de produtos. O produto da TV aberta são os programas. Esses programas podem ter autoria própria ou ter a licença contratada de outra empresa. Podemos classificá-los em 2 categorias, nacionais e internacionais. No âmbito nacional o poder de barganha dos fornecedores é baixo, pois o público alvo se encontra na mídia tradicional e no Brasil existem poucas opções e uma discrepância grande de audiência. Os internacionais possuem um poder

de barganha maior pois além de atenderem ao mundo todo, também atendem diferentes tipos de mídia e plataformas.

Poder de negociação dos clientes: Cliente é quem gera rentabilidade para a empresa. No caso das televisões abertas, quem gera essa rentabilidade são os anunciantes. A televisão atrai a atenção dos telespectadores com os programas e durante intervalos propaga os anúncios. Na atualidade, devido a globalização e ao acesso facilitado a fontes de informação e entretenimento, existem diversas alternativas para empresas que desejam propagar seus produtos, que vão de anúncios em sites e vídeos. Também existem produtos digitais que não possuem em seu modelo de negócio o conceito de anunciantes. Apesar disso, pesquisas indicam que as pessoas preferem assistir anúncios e pagar mais barato ao invés de pagar mais caro por um produto sem anúncios, o que pode acarretar uma mudança de proposta dos substitutos. Em suma, existe uma competitividade quente no mercado e os clientes possuem uma gama de opções muito grande.

Ameaça de produtos substitutos: É importante primeiro analisar a "big picture" dos produtos oferecidos pela grande mídia. De forma geral, a intenção é gerar entretenimento para a audiência. Com a chegada dos computadores e principalmente da recente "geração smartphone" as pessoas têm um acesso muito maior a diferentes fontes de entretenimento e até mesmo nascem conectadas. Está na ponta do dedo, literalmente. É de se observar a modificação no consumo que mais apetece o público. De um tempo pra cá, a preferência é por vídeos curtos e rápidos. Em 2021, tivemos um boom de um aplicativo que possui essa proposta, ele conseguiu se manter no topo e brigou com gigantes. Isso forçou a adaptação da mídia digital, fazendo com que os concorrentes entrassem nessa tendência. Concluímos que a ameaça é grande, pois além de ser um cardápio vasto, novas tendências podem surgir e mudar o esquema do jogo.

Ameaça de entrada de novos concorrentes: O mercado é um mar vermelho de oportunidades. A entrada de concorrentes na tv aberta é extremamente dificultosa pois as empresas que possuem relevância tem longevidade e procuram a identificação com os consumidores, especialmente filiais regionais, que apelam para a empatia e buscam a fidelidade, com sucesso.

Rivalidade entre os concorrentes: As marcas já estão consolidadas e possuem um público fiel, podendo existir até mesmo uma relação de amor ou ódio entre o público. Há uma diferença muito grande de audiência entre as competidoras, análogo a um monopólio. Entre as concorrentes existe uma tensão sobre o câmbio de funcionários entre os grandes players (Globo x Record, por exemplo), pois são considerados representantes da marca e carregam um poder de publicidade com eles. De um tempo pra cá, houve uma flexibilização dando uma maior liberdade aos artistas, mas ainda não é uma prática comum.

4.1.2. Análise SWOT

MATRIZ SWOT – FOFA		
	Fatores Positivos	**Fatores Negativos**
Fatores Internos	**Forças** - Marca consolidada e com alta reputação no mercado; - Profissionais capacitados; - Conhecimento do segmento; - Representante de uma rede de TV líder no Brasil; - Aquisições de tecnologias que permitam a expansão da emissora. - Audiência e alcance; - Time de Marketing Forte.	**Fraquezas** - Tradicionalismo; - Sistema obsoleto e sem precisão de dados; - Inflexibilidade na programação devido à contratos; - Time de inovação pequeno; - Dependência de outros agentes para análise de audiência.
Fatores Externos	**Oportunidades** - Expansão da programação; - Espaço exclusivo para comerciais de empresas externas na emissora; - Liderança local entre as emissoras concorrentes; - Exclusividade na trasmissão de grandes eventos (copa do mundo, brasileirão, olimpíadas); - Criação de uma tecnologia de predição de audiência para a emissora criar novas estratégias visando alcançar um maior público.	**Ameaças** - Serviços de streaming virtuais; - Redes sociais, smartphones e internet; - Aderência dos telespectadores aos programas do catálogo que se alteram por períodos de tempo; - Maturidade do mapeamento da programação dos concorrentes.

Figura 1 - Matriz SWOT da TV Gazeta (Fonte: Criação própria).

4.1.3. Planejamento Geral da Solução

a) quais os dados disponíveis (fonte e conteúdo - exemplo: dados da área de Compras da empresa descrevendo seus fornecedores)

Os dados disponíveis apresentam fontes do Kantar Ibope Media(Kantar IBOPE Media -), proveniente do Kantar Media, líder no mercado de pesquisa de mídia da América Latina, que disponibiliza dados para a tomada de decisão de clientes. Em relação ao conteúdo, este contém informações que informam ao cliente o valor da audiência(Rat%), fidelidade(Fid%), "share"(Shr%) e "reach"(Rch%) dentro de um determinado período (data e hora de início) para determinados públicos (faixa etária, classe social e sexo).

b) qual a solução proposta (pode ser um resumo do texto da seção 2.2)

A solução a ser desenvolvida se baseia na alimentação de sistemas de machine learning, a partir de dados do IBOPE, utilizando recursos de modelagem preditiva para medir com

precisão a estimativa de audiência da faixa-horária do canal comparando com score passados, assim como a composição do público espectador que acompanha a programação exibida e suas preferências previstas. A implementação de tal sistema possibilita o retorno de padrões e peso de atributos existentes nos dados tabulados com o score da audiência, de modo que sirva como auxílio nas decisões das produções e cronogramas de programas. Este é um método algoritmo capaz de usar dados inputados e prever o alcance do público potencial em diferentes cenários de distribuição utilizando variáveis, o qual considera o gênero do programa, data semanal, tempo da transmissão (faixa-horária), sexo e classe social, fornecendo padrões de aumento e queda de conteúdos.

A solução é treinável e melhora iterativamente a fim de gerar novas métricas de desempenho com visão ampla do atendimento de necessidades e das decisões estratégicas a serem tomadas para maximizar o retorno de investimentos em programação. Desse modo, é possível reagir com maior flexibilidade e exatidão às mudanças imprevistas, desempenhando ações antecipadas ao evento, quanto à estreia de programas, análise de resultados e a tração de planos de reversão de programas que não tiveram uma audiência tão boa, além de entender quais variáveis estão influenciando com que o produto esteja indo bem ou ruim e o que garante o alcance e a fidelidade do público.

c) qual o tipo de tarefa (regressão ou classificação)

Considerando a necessidade do projeto, de estimar possíveis valores de audiência a partir de uma ou múltiplas entradas(inputs), e os pesos de cada variável nesta predição, é reconhecida a demanda da implementação de métodos de regressão linear. Isso se dá pelo fato de que os valores resultantes esperados, como score de audiência, peso de atributos e outros, devem ser de característica contínua, numérica. Além disso, para inferir a influência destes atributos na predição final, é preciso entender a relação entre variáveis independentes e a variável de saída (output). Desta forma, é possível explicar a preferência pelo modelo regressivo, o qual tem por característica e finalidade, as próprias exigências mencionadas acima.

d) como a solução proposta deverá ser utilizada

O sistema preditivo desenvolvido deve ser utilizado para gerar métricas de dados acerca do desempenho da audiência, levando em consideração as variáveis desejáveis delimitadas preliminarmente, para assim, possibilitar a criação de estratégias internas que atendam as necessidades da empresa.

Esse sistema deverá ser continuamente alimentado com dados do IBOPE para realização da conversão desses dados para as métricas de estimativa de audiência das faixas-horárias do canal, uma vez que o modelo preditivo é treinável e necessita de inputs de dados para obtenção de uma maior maturidade da predição fornecida. O acesso do sistema por usuários será feito via plataforma Google Colab, a qual apresentará uma organização visual mostrando a segmentação de variáveis que consideram diferentes cenários. Por meio de campos de formulários que utilizarão filtros, o usuário terá acesso a uma estrutura de controle para selecionar e executar opções que irão coletar informações do sistema. Os resultados dos parâmetros gerados serão mostrados em formatos de gráficos.

Posteriormente, o usuário poderá utilizar os resultados obtidos através do sistema de predição para elaborar estratégias que visem impactar a programação da emissora e a venda de espaços publicitários para empresas externas, corroborando para um aumento significativo de lucros e alcance da marca.

e) quais os benefícios trazidos pela solução proposta

- análise de diferentes perspectivas (em relação ao desempenho de programas e
- emissoras);
- visão mais ampla de como a audiência se comporta entre os diferentes canais;
- exploração dos principais fatores que afetam a audiência;
- diminuição de implicações financeiras e maior retorno de investimentos;
- estimativa instantânea dos resultados;
- melhor seleção de conteúdos considerados atraentes pelo público;
- elaboração de estratégias que visam o aumento da popularidade;
- decisão baseada em dados mais precisos.

f) qual será o critério de sucesso e qual medida será utilizada para o avaliar

O desempenho do sistema preditivo poderá ser avaliado através do modelo desenvolvido que apresentará a emissora diferentes cenários possíveis para sua audiência baseado em métricas definidas que permitirão criação de estratégias assertivas para impulsionamento da marca. Além disso, com relação a venda de espaços publicitários ao longo da programação, o sistema poderá gerar um maior retorno financeiro para a emissora, visto a possibilidade de criação de nichos de alcance de audiência para públicos específicos, atendendo o objetivo de comerciais de diferentes produtos ou serviços.

4.1.4. Value Proposition Canvas

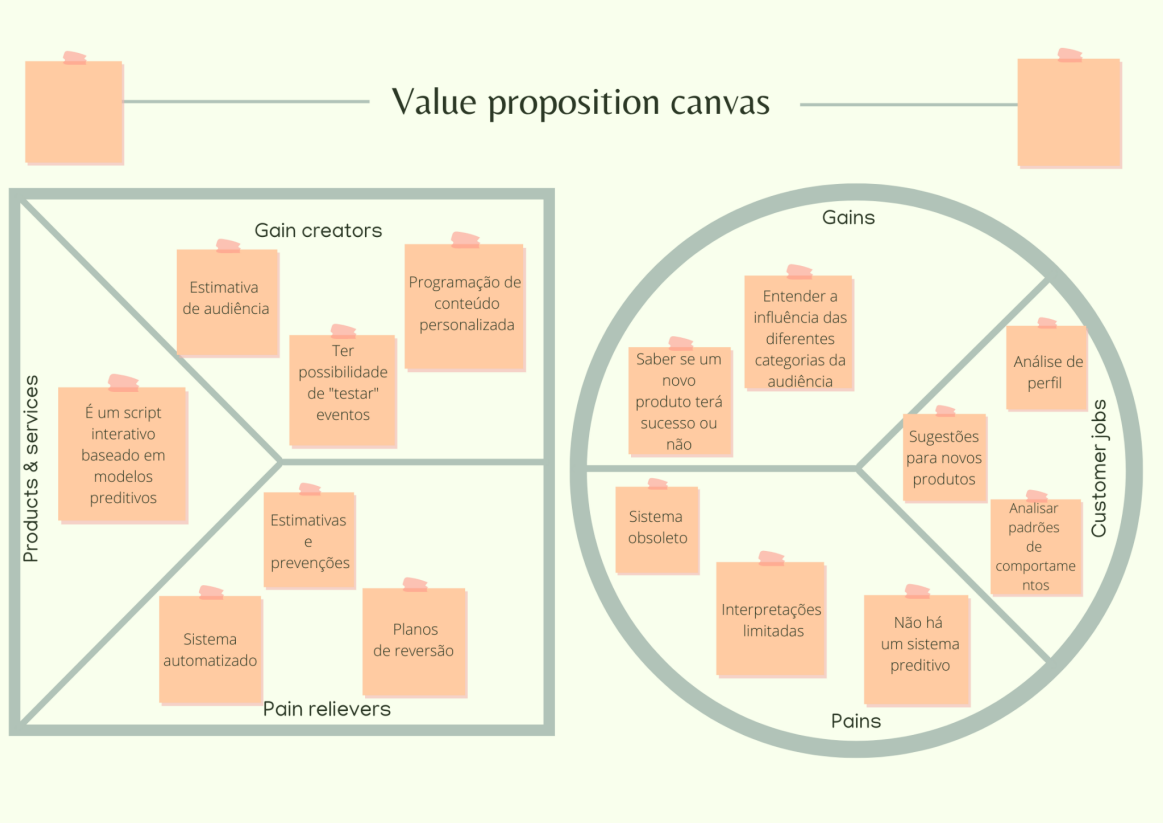


Figura 2 - Value proposition canvas (Fonte: Criação própria).

4.1.5. Matriz de Riscos

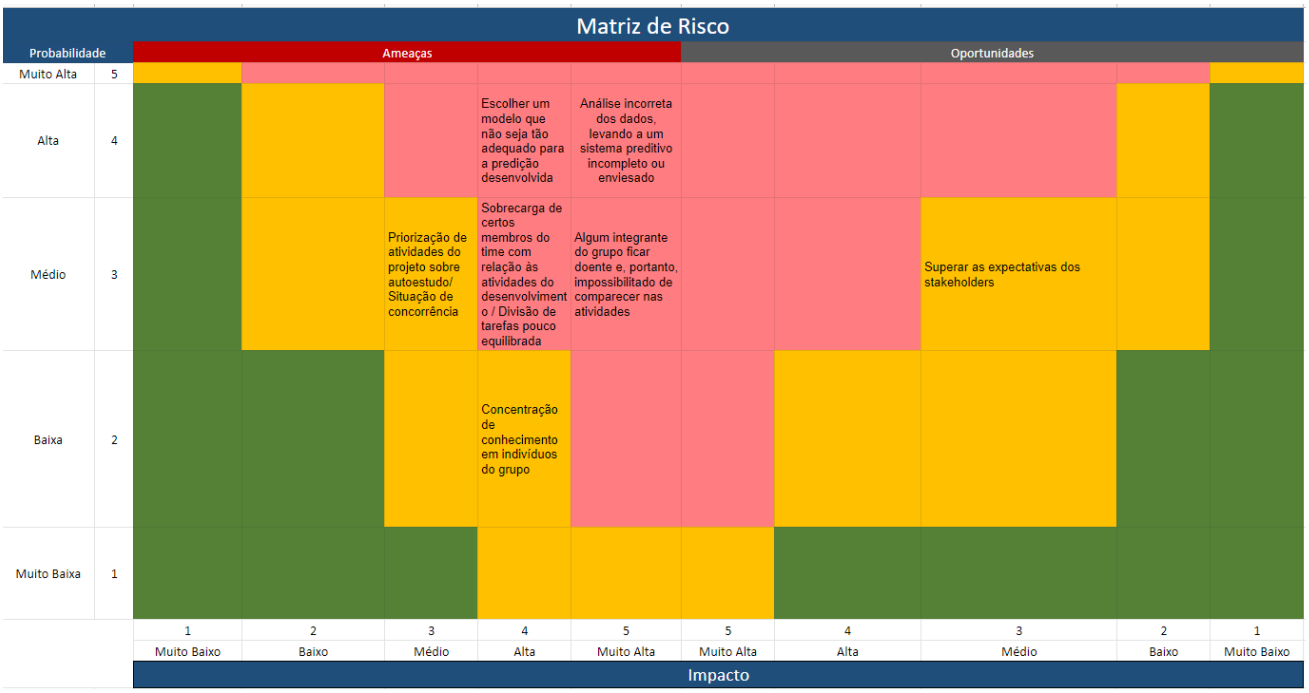


Figura 2 - Matriz de risco desenvolvida pelo grupo (Fonte: Criação própria).

4.1.6. Personas

Persona - *Representante da TV Gazeta*

Nome: Thiago Silva Schneider

"Doravante, de agora em diante; em direção ao futuro."

Idade: 37 anos

Ocupação: Gerente de Operação e Programação

Biografia:

- Nasceu em Vitória, ES;
- Graduado em Engenharia da Computação pela UFES;
- Mestrado em Business Analytics pela UFRGS;
- 5 anos de atuação no mercado de Data science;
- Com 7 anos de experiência na TV Gazeta com programação e operações, começou como coordenador de programação de TV e hoje exerce o cargo de gerente de programação e operações.

Interesse:

- Ter a possibilidade de ter um menor custo de gastos com armazenamento de dados;
- Compreender o impacto e participação de certos atributos na taxa de audiência final;
- Ter uma publicidade mais direcionada a públicos e nichos específicos.

Motivações:

- Precisão de alcance de audiência de um novo programa;
- Atrair um maior público para a emissora;
- O processo que é atualmente feito é impreciso e manual.

Dores:

- Alto Gasto com armazenamento de dados;
- Imprevisibilidade do sistema*;
- Ausência de um sistema automatizado;
- Falta de assertividade do sistema;
- Falta de agilidade na geração de resultados

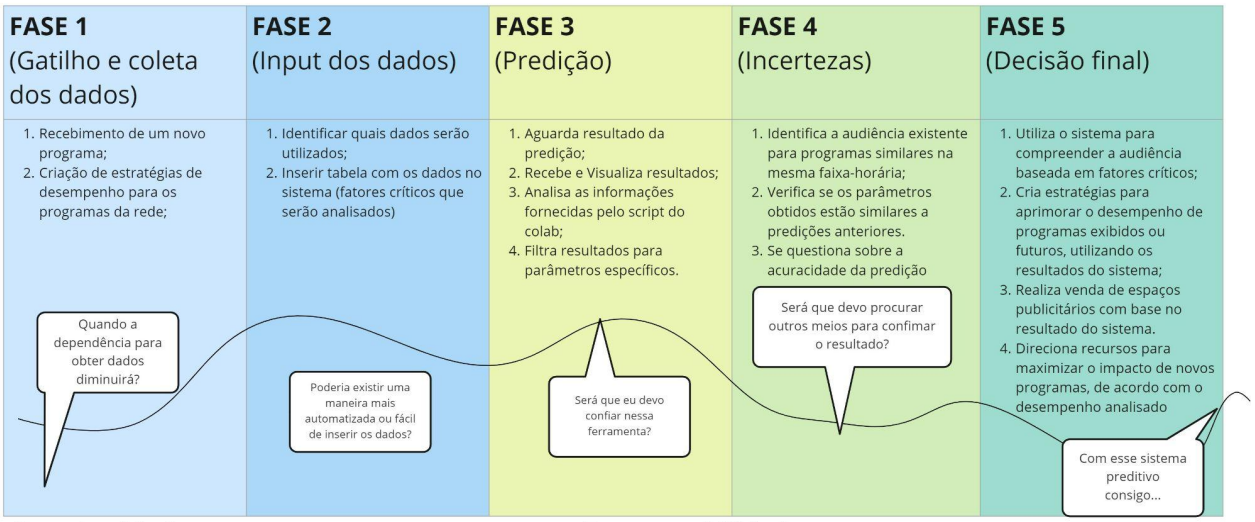
4.1.7. Jornadas do Usuário

Thiago Silva Schneider

Cenário: Thiago precisa compreender fatores críticos do público em determinada faixa horária para inserção de um evento, seja ele um novo programa ou comercial.

Expectativas

Espera-se que o tempo excessivo na preparação de dados seja utilizado em insights, obtendo assertividade na predição da taxa de audiência para novas programações e redução de gastos.



Oportunidades

- Estruturação de um time de datascience, para a manutenção e atualizações do modelo, a fim de torná-lo mais eficiente e atual.
- Criação de interface Web para facilitar o uso do script
- Comparar a audiência estimada com outras emissoras para um mesmo horário

Responsabilidades

Atribua responsabilidades a pessoas/equipes e respectivas ações para resolver problemas e alcançar as melhorias

- Responsabilidades do time de marketing de gerenciar os recursos publicitários frente aos dados obtidos pela ferramenta
- RH para a contratação de novos desenvolvedores e cientistas de dados
- Responsabilidade dos cientistas de dados de constantemente alimentarem o modelo
- Equipe de UX para desenvolver o conceito de uma nova plataforma para a ferramenta.

Figura 3 - Jornada do usuário (Fonte: Criação própria)

4.2. Compreensão dos Dados

1. **Descreva os dados a serem utilizados (disponibilizados pelo cliente e outros se tiverem sido incluídos), detalhando a fonte, o formato (CSV, XLSX, banco de dados, etc.), o conteúdo e o tamanho.**

A fonte dos dados disponíveis vem do Kantar Ibope Media(Kantar IBOPE Media -), proveniente do Kantar Media. Enquanto o formato do arquivo dos dados é XLSX, ou seja, Planilha do Microsoft Office Excel.

Além disso, o conteúdo desse arquivo contém 18 abas, cada aba referencia uma emissora e um período da semana. Sendo 3 delas referenciando a emissora Total-Ligados-Especial("TLE - Seg a Sex", "TLE - Sáb" e "TLE - Dom"), outras 3 referenciando a emissora Principal, do parceiro("Emissora Principal (xxx) - Seg a Sex", "Emissora Principal (xxx) - Sab" e "Emissora Principal (xxx) - Dom"), mais 3 referenciam a emissora concorrente A ("Emissora concorrente A (yyy) - Seg a Sex", "Emissora Concorrente A(yyy) - Sab" e "Emissora concorrente A (xxx) - Dom"), mais 3 referenciando a emissora concorrente B("Emissora concorrente B (zzz) - Seg a Sex", "Emissora concorrente B (zzz) - Sab" e "Emissora concorrente B (zzz) - Dom"), restando 6 abas, incluindo Canais Pagos("Canais Pagos(OCP) - Seg a Sex", "Canais Pagos(OCP) - Sab", "Canais Pagos(OCP) - Dom") e NI conteúdo, serviços de streaming que não são canais ou emissoras da televisão("NI Conteúdo - Seg a Sex", "NI Conteúdo - Sab", "NI Conteúdo - Dom"). Enquanto nas colunas, estão inclusas: "Dia"(XX/XX/XXXX), "Hora de Início", "Emissora", "Dia da Semana", várias delas relacionam o "Rat"(valor de audiência), "Shr%"(Share), "Rch%"(Reach), "Fid%"(Fidelidade), usando de base perfis de público(AB, C1, C2, DE, Masculino e Feminino) e diversos intervalos de faixa etária para avaliar esses 4 parâmetros.

O tamanho do arquivo que contém todos esses dados é de 437.754 KB.

Nome do atributo	Tipo do atributo	Descrição
Emissora	Texto	Empresa produtora e transmissora dos conteúdos daquela aba da tabela
Hora Início	Tempo	Horário que inicia a medição de audiência em uma determinada data
Data	Data	Data referência daquela medição de uma determinada emissora
Rat%(Rating)	Real	Valor da audiência dos programas
Shr%(Share)	Real	

Nome do atributo	Tipo do atributo	Descrição
Rch%(Reach)	Real	É o RAT e divide pelo TLE
Fid%(Fidelidade)	Real	É a fidelidade do público em relação a emissora
Faixa-Etária	Texto	Intervalo de idade do público telespectadores
Masculino	Texto	Intervalo que contém strings com a características
Feminino	Texto	Intervalo que contém strings com a características
AB	Texto	Intervalo da classe social A - B do público, associado a Rating, Share, Reach e Fidelidade
C1	Texto	Intervalo da classe social C do público associado a Rating, Share, Reach e Fidelidade
C2	Texto	Intervalo da classe social C2 do público associado a Rating, Share, Reach e Fidelidade
DE	Texto	Intervalo da classe social D - E do público associado a Rating, Share, Reach e Fidelidade
4-11 anos	Real	Intervalo de idade do público, associado aos atributos Rating, Share, Reach e Fidelidade
12-17 anos	Real	Intervalo de idade do público, associado aos atributos Rating, Share, Reach e Fidelidade
18-24 anos	Real	Intervalo de idade do público, associado aos atributos Rating, Share, Reach e Fidelidade
25-34 anos	Real	Intervalo de idade do público, associado aos atributos Rating, Share, Reach e Fidelidade
35-49 anos	Real	Intervalo de idade do público, associado aos atributos Rating, Share, Reach e Fidelidade
50-59 anos	Real	Intervalo de idade do público, associado aos atributos Rating, Share, Reach e Fidelidade
69+ anos	Real	Intervalo de idade do público, associado aos atributos Rating, Share, Reach e Fidelidade

Tabela 1 - Interpretação da descrição dos dados a serem utilizados

(Fonte: Planilha disponibilizada pelo parceiro de mercado).

a. **Se houver mais de um conjunto de dados, descrição de como serão agregados/mesclados.**

A organização dos dados foi realizada individualmente para as tabelas da Emissora Principal e Emissora Concorrente B, nas quais foram mescladas informações de todos os dias e finais de semana em uma única coluna, ordenados por data, horário de início e os respectivos programas e categorias da faixa horária disposta. Após isso, foram agregados os dados da Grade Diária em cada uma das tabelas. Ademais, foi-se adicionando uma coluna com o nome do mês e dia do mês para diferenciação na segmentação dos dados.

b. **Descrição dos riscos e contingências relacionados a esses dados (qualidade, cobertura/diversidade e acesso).**

- **Qualidade**

Para considerar dados que possuem qualidade, partimos de três pilares: a integridade, a acuracidade e a completude. A integridade indica a segurança dos dados contidos na fonte. A acuracidade indica quanto os dados da fonte representam a realidade. A completude indica quanto de todos os dados necessários para atender a demanda do negócio está presente na fonte. Partindo desse pressuposto, o fato de não haver dados sobre identidade de gênero, raça, etnia e orientação, fazem com que sejam incompletos. Dados de qualidade eliminam problemas relativos ao negócio da organização como perda de receita, altos custos de produção, incapacidade de manter seus clientes fiéis, perda de mercado, dentre outros.

Os dados disponibilizados são limitados e não abrangem as informações citadas anteriormente, o que dificulta a obtenção de uma visão panorâmica, tal como extrair mais dados de pessoas que possuem os aparelhos em seus lares, já que os dados têm uma função fundamental na implementação de programas voltados para diversidade e inclusão como programações mais específicas que abrangem uma maior parcela da população, selecionando conteúdo para nichos específicos que interessam ao público.

Além disso, a pesquisa é feita por amostragem e não envolve toda a população da região e, por não ter muitos aparelhos instalados, os dados muitas vezes não são tão fiéis à realidade.

- **Diversidade**

Os dados abrangem pessoas do sexo feminino e masculino divididas em grupo nas seguintes faixa-etária:

➔ 60+ anos
➔ 50 a 59 anos
➔ 35 a 49 anos
➔ 25 a 34 anos
➔ 18 a 24 anos
➔ 12 a 17 anos

➔ 4 a 11 anos

- **A exibição da programação está dividida em:**

➔ Data;
➔ Hora de início;
➔ Emissora;
➔ Dia da semana;
➔ Quantidade de domicílios que assistem em rat% e share%;
➔ Classe do espectador dividida em AB, C1, C2

- **Acesso**

O aparelho chamado PeopleMeter, a medida que identifica o canal que o espectador está acompanhando, envia informações diariamente para a central através do sinal de radiofrequência. Os dados referente a audiência são disponibilizados através do Instituto Brasileiro de Opinião Pública e Estatística, o IBOPE, que posteriormente são tabulados e repassados para as emissoras de TV que possuem acesso aos resultados.

c. Se aplicável: descrição de como será selecionado o subconjunto para análises iniciais (quando o tamanho do conjunto de dados impossibilita a utilização do conjunto completo em todas as etapas da definição do modelo a ser usado).

A análise inicial será limitada a emissora da empresa parceira e uma concorrente direta. O subconjunto selecionado será a média dos atributos da tabela por sessões de uma hora para cada dia de uma única semana do mês durante o período de 24 meses (2 anos).

d. Se houver: descrição das restrições de segurança.

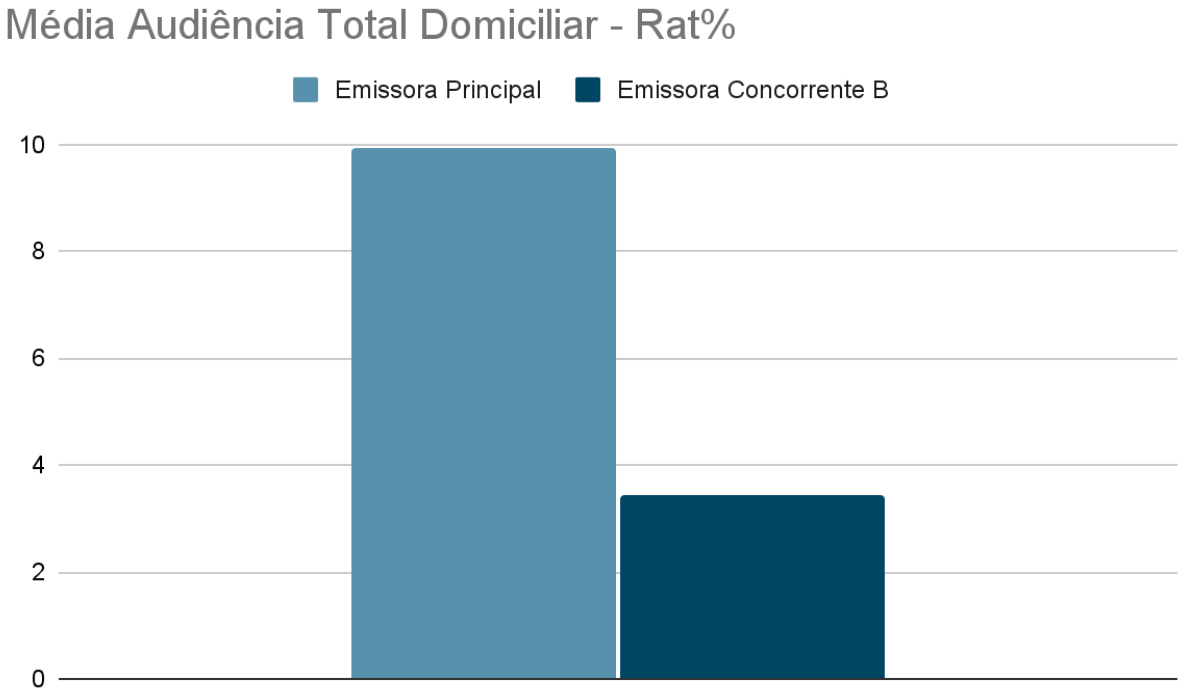
A Rede Gazeta e a Kantar Ibope possuem um contrato de confidencialidade que veta a divulgação do nome das emissoras. Portanto, usaremos nomes fantasias (exemplos: Concorrente 1, Concorrente 2 e etc).

2. Descrição estatística básica dos dados, principalmente dos atributos de interesse, com inclusão de visualizações gráficas e como essas análises embasam suas hipóteses.

Análise Descritiva:

Dados: Audiência por Total Domiciliar(Rat%) Emissora Principal x Emissora concorrente B

Métrica	Emissora Principal	Emissora Concorrente B
Média Audiência Total Domiciliar %	9.92	3.45
Moda Audiência Total Domiciliar %	3.72	0.0
Mediana Audiência Total Domiciliar %	8.41	2.87
Máximo Audiência Total Domiciliar %	45.34	37.7
Mínimo Audiência Total Domiciliar %	0.0	0.0



Comparando os dados anteriores, é notória a percepção de 6.47 pontos de diferença, na audiência média aferida, entre a Emissora Principal e a Emissora Concorrente B. Isso demonstra, de forma nítida, a discrepância entre o índice de audiência no que se refere aos telespectadores, das duas emissoras comparadas, presente até nos mais altos picos de medida.

Dados: Audiência por Gênero (Rat%) -Emissora Principal x Emissora concorrente B

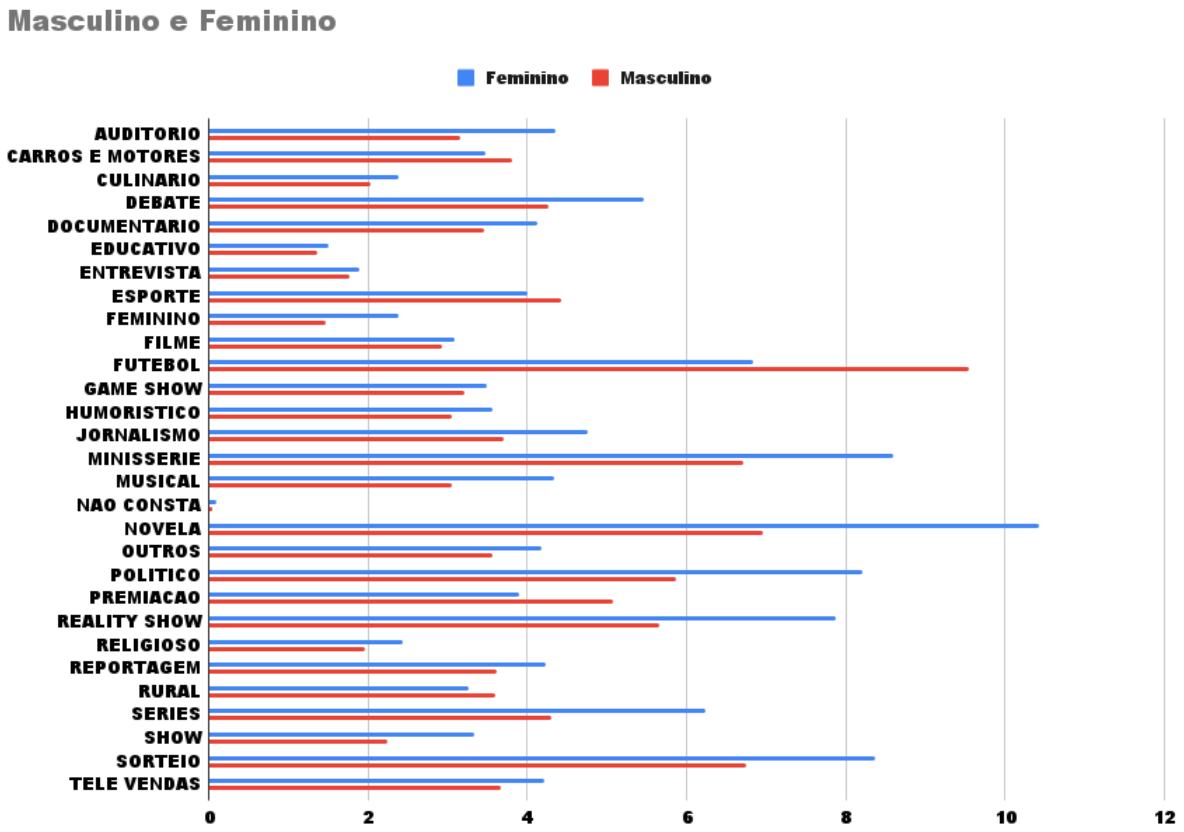
Métrica	Emissora Principal	Emissora Concorrente B
Média de Audiência por Gênero Feminino (Rat%)	5.14	1.84
Média de Audiência por Gênero Masculino (Rat%)	4.03	1.37

Moda Audiência por Gênero Feminino (Rat%)	0.0	0.0
Moda Audiência por Gênero Masculino (Rat%)	0.0	0.0
Mediana Audiência por Gênero Feminino (Rat%)	4.08	1.54
Mediana Audiência por Gênero Masculino (Rat%)	3.41	1.14
Máximo Audiência por Gênero Feminino (Rat%)	24.40	16.59
Máximo Audiência por Gênero Masculino (Rat%)	24.09	20.68
Mínimo Audiência por Gênero Feminino (Rat%)	0.0	0.0
Mínimo Audiência por Gênero Masculino (Rat%)	0.0	0.0

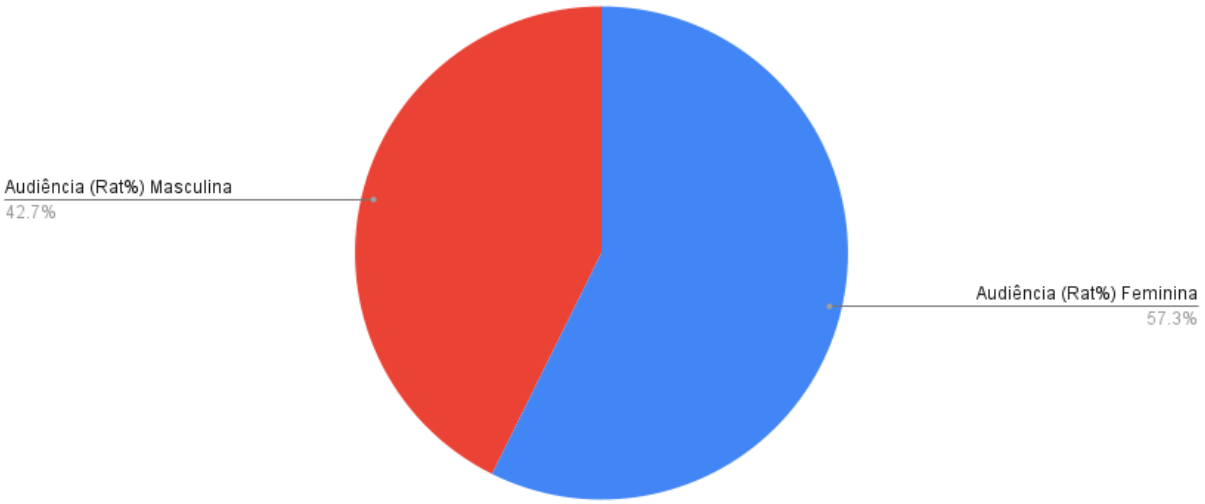
A partir do cálculo das médias das audiências masculinas e femininas, chegou-se à conclusão de que na Emissora Principal, é predominante a presença do público feminino, compondo aproximadamente 56.1% do público total. Para este cálculo, foram comparadas as médias dos valores Rat% de cada gênero nos últimos dois anos. Fazendo uma comparação direta entre elas, chegou-se neste gráfico abaixo.

Porcentagem do Público Feminino e Masculino - Emissora Principal:

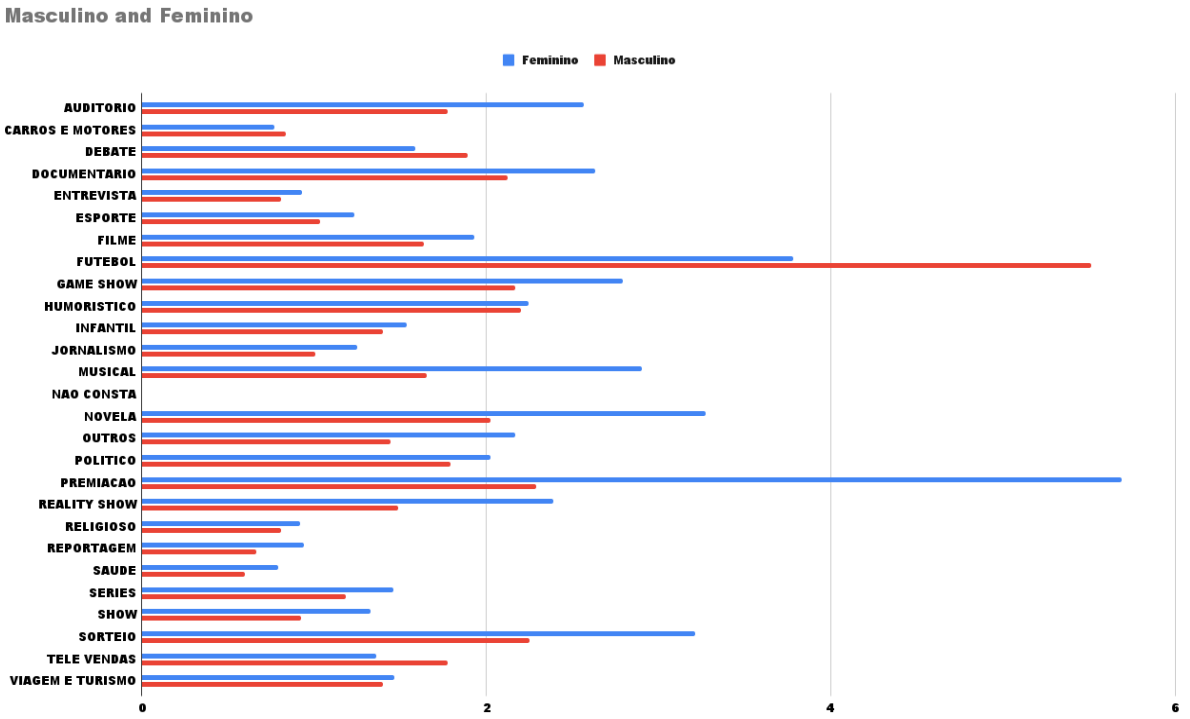
Público Feminino e Masculino em Relação às Categoria - Emissora Principal



Porcentagem do Público Feminino e Masculino - Emissora Concorrente B



Público Feminino e Masculino em Relação às Categoria - Emissora Concorrente B

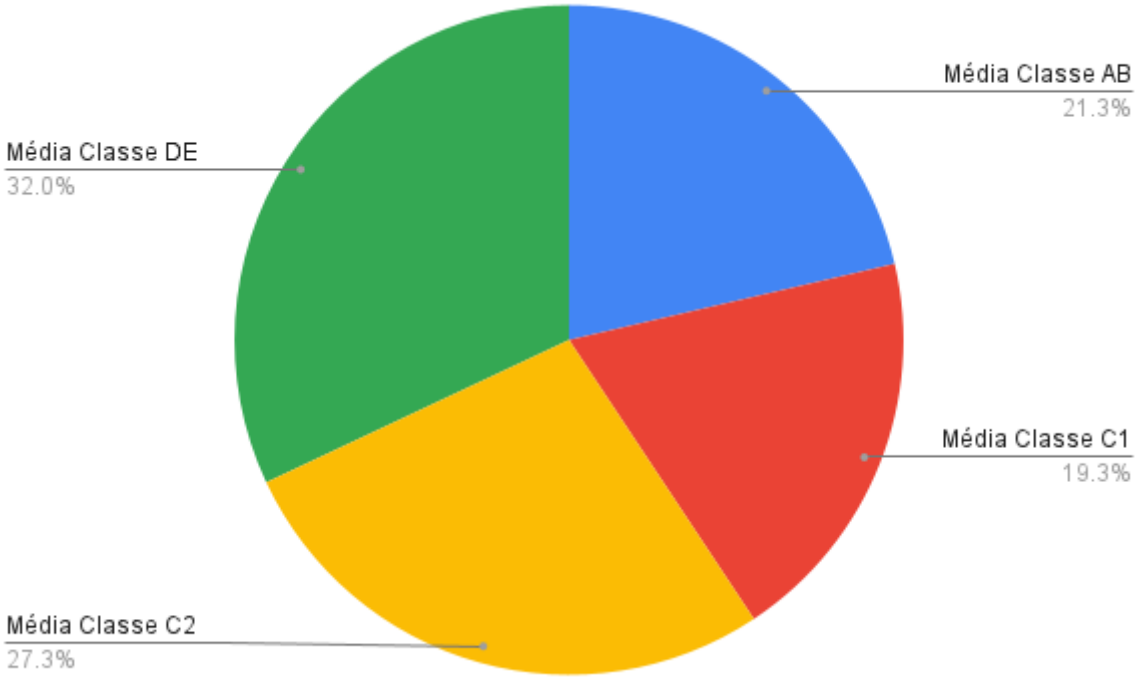


Dados: Audiência por Classe Social (Rat%) Emissora Principal x Emissora concorrente B

Métrica	Emissora Principal	Emissora Concorrente
Média de Audiência por Classe Social AB (Rat%)	4.03	0.89
Média de Audiência por Classe Social C1 (Rat%)	3.65	1.20
Média de Audiência por Classe Social C2 (Rat%)	5.16	1.34
Média de Audiência por Classe Social DE (Rat%)	6.04	3.87
Moda Audiência Classe Social AB (Rat%)	0.0	0.0
Moda Audiência Classe Social C1 (Rat%)	0.0	0.0

Moda Audiência Classe Social C2 (Rat%)	0.0	0.0
Moda Audiência Classe Social DE (Rat%)	0.0	0.0
Mediana Audiência Classe Social AB (Rat%)	3.51	0.58
Mediana Audiência Classe Social C1 (Rat%)	2.96	0.76
Mediana Audiência Classe Social C2 (Rat%)	3.95	0.85
Mediana Audiência Classe Social DE (Rat%)	4.70	3.43
Máximo Audiência Classe Social AB (Rat%)	22.28	16.31
Máximo Audiência Classe Social C1 (Rat%)	22.40	23.78
Máximo Audiência Classe Social C2 (Rat%)	29.83	17.44
Máximo Audiência Classe Social DE (Rat%)	31.44	25.96
Mínimo Audiência Classe Social AB (Rat%)	0.0	0.0
Mínimo Audiência Classe Social C1 (Rat%)	0.0	0.0
Mínimo Audiência Classe Social C2 (Rat%)	0.0	0.0
Mínimo Audiência Classe Social DE (Rat%)	0.0	0.0

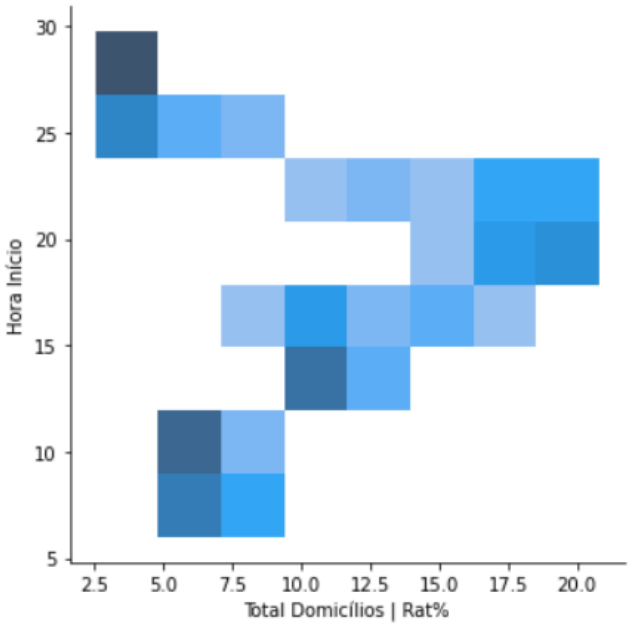
Público Dividido em Relação às Classe Sociais - Emissora Principal:



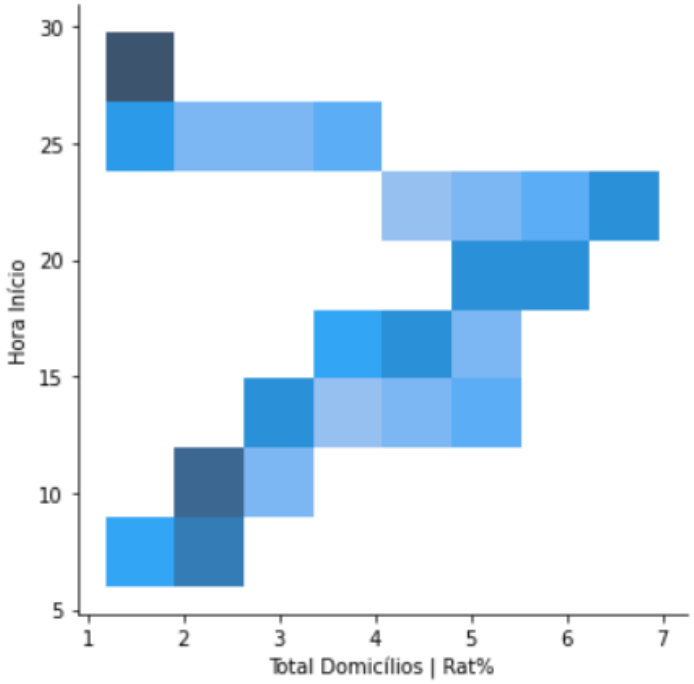
Referente ao seccionamento de classes, e sua distribuição na audiência total da Emissora Principal, verifica-se que a classe D e E possuem maior participação (32.0%), seguida pela classe C2 (27.3%), classes A e B (21.3%), em último lugar, a classe C1 (19.3%). Percebe-se que, em geral, há uma proximidade muito grande entre as médias das classes A e B, e a da classe C1, assim como a das classes D e E e as classes C2. Isso, de certa forma, reflete a realidade Brasileira atual, onde, segundo um levantamento feito pela Consultoria Tendências, é revelado que as classes D e E são mais da metade das casas no Brasil.

Dados: Concentração de Taxas de Audiência por Hora de Início (Rat%) Mapa de Calor

Mapa de Calor - Emissora Principal



Mapa de Calor - Emissora Concorrente B



Comparando a audiência total, por hora de início, percebe-se uma grande concentração de altas taxas de audiência Total Domiciliar, no período próximo entre 18h e 20h. Ao comparar a grade de eventos da Emissora Principal, nota-se uma grande presença de episódios de novela, intermitentes nesta faixa. Logo em seguida, das 20h até as 22h, há uma leve queda nesta medida, mas permanece a alta concentração de audiência. Extraindo a grade desta última faixa, encontra-se por quase todo o momento, o programa Jornal Nacional. Além disso, é também perceptível que, no mesmo período de alta, a emissora concorrente B apresenta uma queda considerável em sua concentração.

Dados: Audiência por Faixa Etária(Rat%) Emissora Principal x Emissora concorrente B

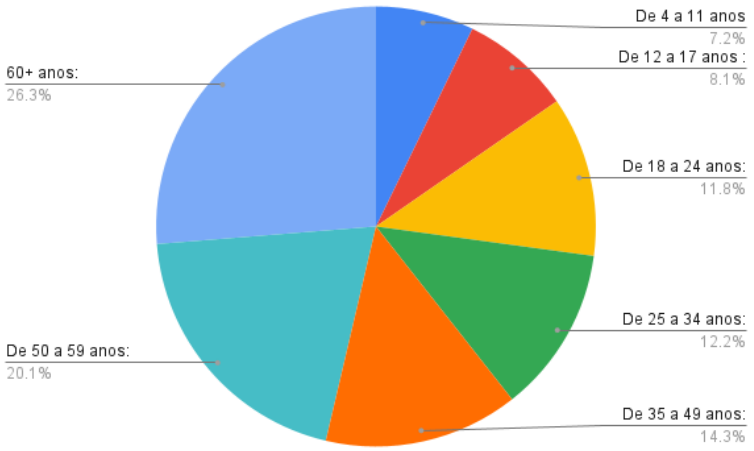
Métrica	Emissora Principal	Emissora Concorrente
Média de Audiência por Faixa Etária 4 - 11 anos (Rat%)	2.2	1.68
Média de Audiência por Faixa Etária 12 - 17 anos (Rat%)	2.51	1.32
Média de Audiência por Faixa Etária 18 - 24 anos (Rat%)	3.63	1.07
Média de Audiência por Faixa Etária 25 - 34 anos (Rat%)	3.77	1.31
Média de Audiência por Faixa Etária 35 - 49 anos (Rat%)	4.40	1.31
Média de Audiência por Faixa Etária 50 - 59 (Rat%)	6.18	1.94
Média de Audiência por Faixa Etária 60+ anos (Rat%)	8.10	2.60
Moda de Audiência por Faixa Etária 4 - 11 anos (Rat%)	0.0	0.0
Moda de Audiência por Faixa Etária 12 - 17 anos (Rat%)	0.0	0.0
Moda de Audiência por Faixa Etária 18 - 24 anos (Rat%)	0.0	0.0
Moda de Audiência por Faixa Etária 25 - 34 anos (Rat%)	0.0	0.0
Moda de Audiência por Faixa Etária 35 - 49 anos (Rat%)	0.0	0.0
Moda de Audiência por Faixa Etária 50 - 59 anos (Rat%)	0.0	0.0
Moda de Audiência por Faixa Etária 60+ anos (Rat%)	0.0	0.0
Mediana Audiência por Faixa Etária 4 - 11 anos (Rat%)	1.47	1.04

Mediana de Audiência por Faixa Etária 12 - 17 anos (Rat%)	1.57	0.22
Mediana de Audiência por Faixa Etária 18 - 24 anos (Rat%)	2.61	0.00
Mediana de Audiência por Faixa Etária 25 - 34 anos (Rat%)	2.89	0.98
Mediana de Audiência por Faixa Etária 35 - 49 anos (Rat%)	3.78	0.94
Mediana de Audiência por Faixa Etária 50 - 59 anos (Rat%)	5.01	1.49
Mediana de Audiência por Faixa Etária 60+ anos (Rat%)	6.68	1.81
Máximo de Audiência por Faixa Etária 4 - 11 anos (Rat%)	20.25	16.82
Máximo de Audiência por Faixa Etária 12 - 17 anos (Rat%)	22.72	18.57
Máximo de Audiência por Faixa Etária 18 - 24 anos (Rat%)	28.07	16.68
Máximo de Audiência por Faixa Etária 25 - 34 anos (Rat%)	30.15	28.73
Máximo de Audiência por Faixa Etária 35 - 49 anos (Rat%)	21.90	21.07
Máximo de Audiência por Faixa Etária 50 - 59 anos (Rat%)	32.51	19.29
Máximo de Audiência por Faixa Etária 60+ anos (Rat%)	37.64	24.53
Mínimo de Audiência por Faixa Etária 4 - 11 anos (Rat%)	0.0	0.0
Mínimo de Audiência por Faixa Etária 12 - 17 anos (Rat%)	0.0	0.0
Mínimo de Audiência por Faixa Etária 18 - 24 anos (Rat%)	0.0	0.0
Mínimo de Audiência por Faixa Etária 35 - 49 anos (Rat%)	0.0	0.0

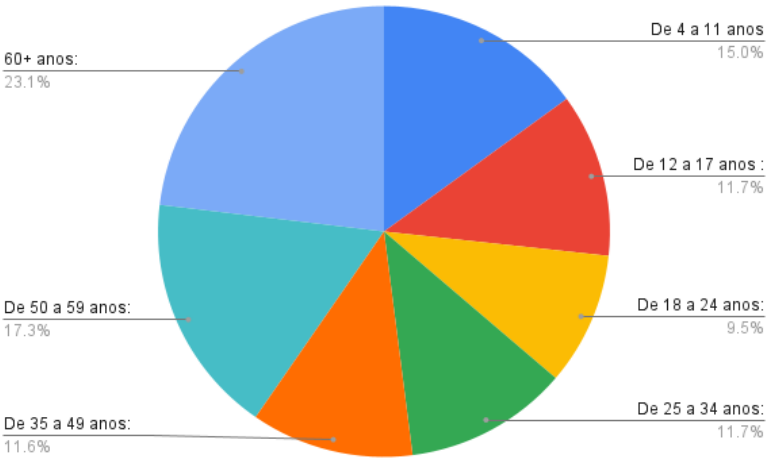
Mínimo de Audiência por Faixa Etária 50 - 59 anos (Rat%)	0.0	0.0
Mínimo de Audiência por Faixa Etária 60+ anos (Rat%)	0.0	0.0

A partir da análise das médias de audiência por faixa etária, infere-se que a maior parcela da audiência da rede, 25.7%, se deve à pessoas com 60 ou mais anos. Em segundo lugar, posicionam-se pessoas entre 50 a 59 anos, com 20.3% da parcela. Seguido por pessoas com 35 a 49 anos(14.3%), 25 a 34 anos (12.3%), 18 a 24 anos (12.0%), 12 a 17 anos (8.3%) e 4 a 11 anos (7.2%). Com isso, conclui-se que a audiência predominante é a da população idosa, seguida pela meia-idade.

Público Dividido em Relação às Faixa Etária - Emissora Principal



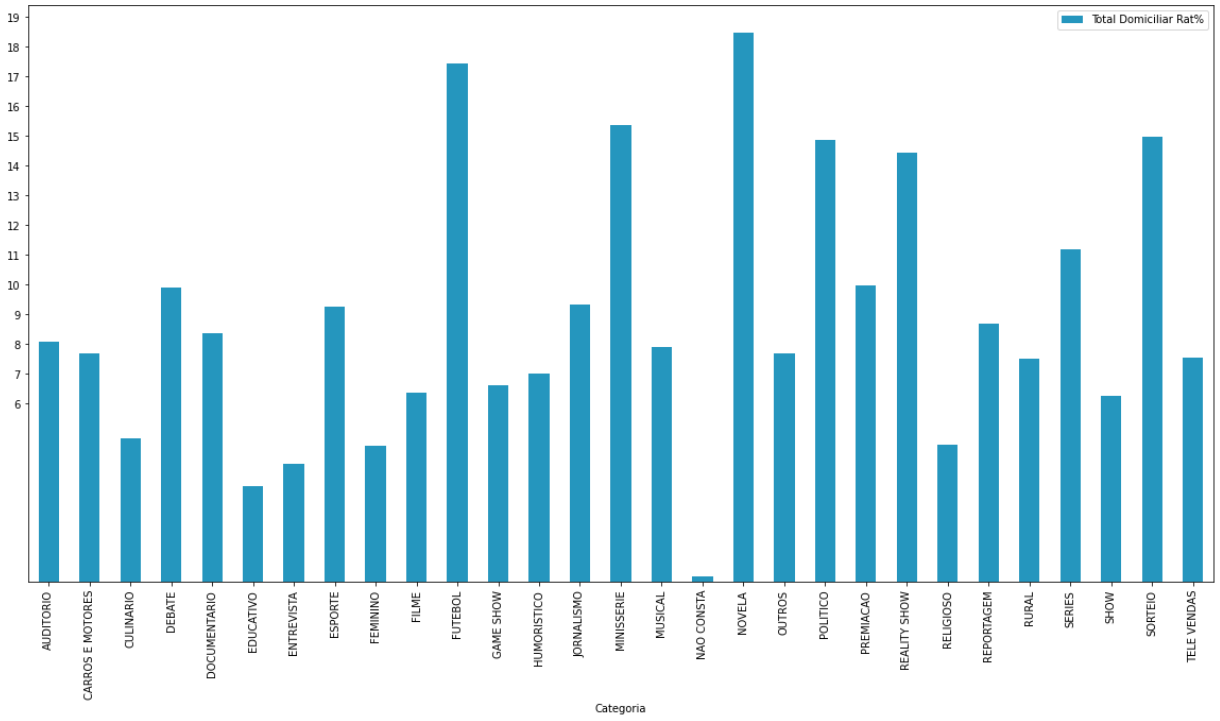
Público Dividido em Relação às Classe Sociais - Emissora Concorrente B



Dados: Audiência por Programação de Conteúdos Categorizados (Rat%) Emissora Principal x Emissora Concorrente B

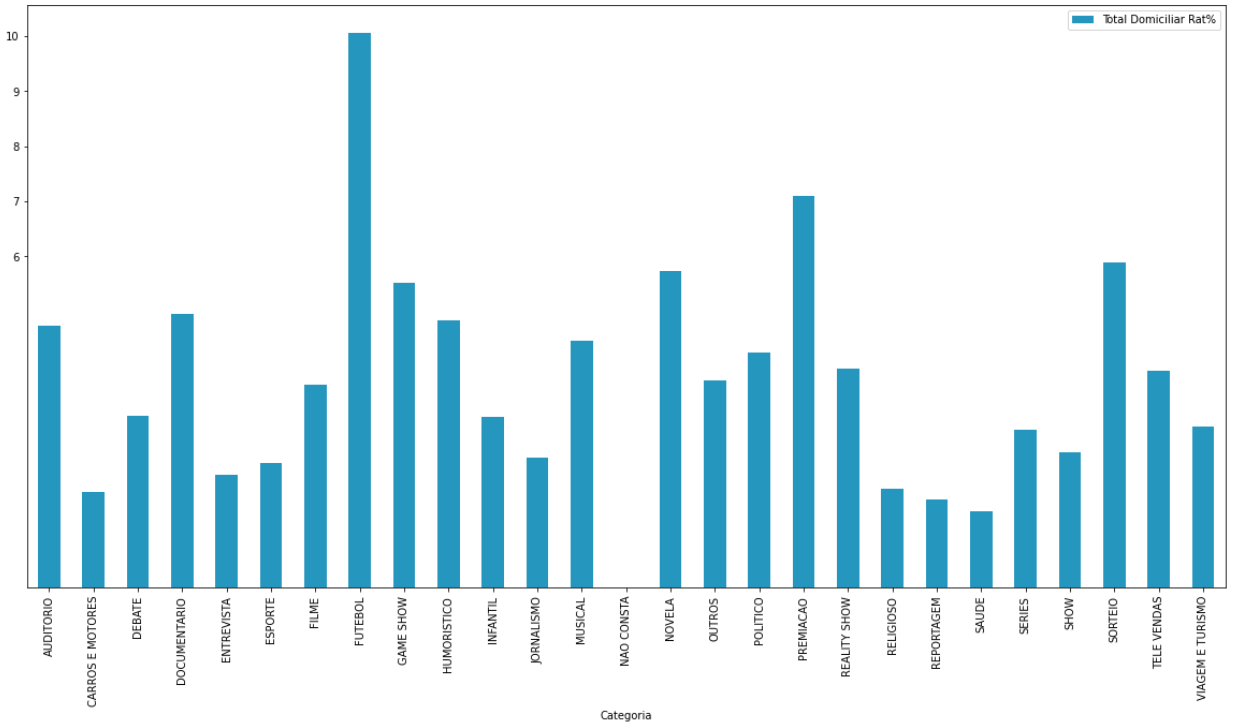
Categoria	Pontos Emissora Principal	Pontos Emissora Concorrente B
Novela	18.49	5.73
Futebol	17.44	10.05
Premiação	9.99	7.08
Sorteio	14.99	5.89
Auditório	8.09	4.75
Tele Vendas	7.53	3.93
Game Show	6.61	5.52
Político	14.86	4.25
Carro e Motores	5.89	1.74

Dados de Audiência em Relação a Conteúdos Categorizados - Emissora Principal



Nesse sentido, há uma relação direta entre as categorias ditas na tabela e a aproximação que o gráfico trouxe. É perceptível, que programas de entretenimento como “FUTEBOL”e “NOVELA” lideram os dados, deixando o “JORNALISMO” em terceiro lugar.

Dados de Audiência em Relação a Conteúdos Categorizados - Emissora Concorrente B



Levando em consideração também, a audiência por categoria, da emissora concorrente B, é visível a alta demanda pela seção “FUTEBOL”, o qual é predominante sobre todas as outras seções. Diferentemente da emissora principal, o conteúdo “NOVELA” se encontra em terceira posição, perdendo lugar para a categoria “PREMIAÇÃO”.

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

A predição utilizará target de dados quantitativos de audiência fornecidos pela coluna de valor de audiência (Rat%). Os dados são compostos por números decimais e o modelo é de natureza contínua, buscando-se estimar o valor mais provável do Rat%. É esperado que ao final da construção sistema preditivo, por meio da aplicabilidade do input destes conjuntos de dados de entrada, a resposta predita apresentada pelo algoritmo seja a mais provável de ocorrer para a data e horário selecionada pelo usuário.

4.3. Preparação dos Dados

4.3.1 Anonimização dos dados

Colab de referência:
https://drive.google.com/file/d/1nrO57J8Xk09h0iDu6xkwPqAh0sSkm5mz/view?usp=sharing

Devido a limitações contratuais do parceiro e com a Kantar Ibope, os nomes das emissoras e quaisquer menções diretas, foram substituídas por referências anonimizadas.

Para isso, efetuamos o seguinte processo:

Substituição dos nomes de emissoras por referências anonimizadas
Célula de Referência: 1.2 Renomeação das colunas sensíveis

Descrição: As colunas da planilha Grade Diária foram alteradas por correspondentes respectivos, anonimizados, transformando-se em “Emissora A”, “Emissora B” e “Emissora C”.

Substituição da programação por referências anonimizadas
Célula de Referência: 1.3 Anonimização dos nomes dos programas

Descrição: Devido à programações específicas de emissoras, foram renomeados os programas por um dicionário respectivo, alterando elementos como “JORNAL ABC” para “programa_n”, sendo “n”, uma representação no dicionário de programas. Isto foi feito utilizando o método de regex, para encontrar e substituir o nome dos programas, sem alterar a estrutura da tabela original. Além disso, a coluna “Praça” foi substituída da mesma forma, por conter dados confidenciais de identificação das emissoras.

Remoção da coluna “Emissora”
Célula de Referência: 2.1 Remoção da coluna Emissora das Tabelas

Descrição: Na planilha de audiência da emissora principal, os arquivos foram renomeados e a coluna “Emissora” foi removida.

4.3.2 Manipulação de dados e registros

Transformação da coluna “Faixa Horária” em “Hora Início"
Célula de Referência: 1.2 Transformação da coluna “Faixa Horária” em “Hora Início”.

Descrição: Divisão do horário por " - "(por meio de um split, que basicamente transforma uma string em lista, permitindo assim a sua divisão em duas ou mais partes) na feature "Faixa Horária". Criamos "Hora Início" como sendo “coluna filha” da coluna "Faixa Horária", atribuindo a ela a padronização de 00:00:00 (hora:minuto:segundo), por meio da concatenação do horário inicial (primeiro elemento extraído do split) + ':00'.

Objetivo: Ter uma padronização uniforme dos dados horários estabelecidos.

Conversão de tipos da coluna “Data”
Célula de Referência: 1.3 Conversão de tipos da coluna “Data”

Descrição: Conversão da coluna “Data” do tipo string para o tipo datetime.
Objetivo: Possibilitar a ordenação correta da coluna “Data” por meio do sort_values (que usaremos na célula 2.4) para considerar a ordem dos valores de datetime.

Derivação dos atributos “Programa” e “Categoria”
Célula de Referência: 1.4 Derivação dos atributos “Programa” e “Categoria”.

Descrição: Divisão da coluna das emissoras para as colunas "Programa" e "Categoria", aplicando isso à emissora “A".
Objetivo: Selecionar as features, para facilitar o filtro das informações que pretendemos obter de acordo com cada emissora desejada.

Concatenação das tabelas dos Dias da Semana - Emissora A
Célula de Referência: 2.2 Concatenação das tabelas dos Dias da Semana - Emissora A.

Descrição: Concatenação (usando o método concat) das tabelas A - Seg a Sex.csv, A - Sab.csv e A - Dom.csv
Objetivo: Concentrar a audiência de todos os dias da semana em uma mesma tabela, possibilitando a comparação assertiva com a organização dos dias da semana na tabela Grade Diária.

Conversão de colunas “Data” para “Datetime”
Célula de Referência: 2.3 Conversão de tabelas de “Data” para “Datetime” **.**

Descrição: Conversão da coluna "Data"' das três tabelas concatenadas, do tipo string para o tipo “datetime”
Objetivo: Fornecer uma padronização para manipulações futuras, como a ordenação correta da coluna “Data” por meio do sort_values, para considerar a ordem dos valores de datetime.

Sort da tabela agregada “emissora_a_geral” pelas colunas “Data” e “Hora Início”
Célula de Referência: 2.4 Ordenamento da tabela agregada “emissora_a_geral” pelas colunas “Data” e “Hora Início”.

Descrição: Ordenação das colunas com base nas colunas "Data" e "Hora Início", por meio do método “sort_values”, que permite a ordenação, já que as colunas foram convertidas de string para datetime.
Objetivo: Ordenar dados da tabela que concatenamos, inicialmente pela ordem das datas e levando em conta a coluna “Hora Início”.

Agregação das duas tabelas manipuladas: “Grade Diária” e “emissora_a_geral”
Célula de Referência: 3.1 Agregação (Merge) das duas tabelas Manipuladas: "Grade Diária" e "Emissora A".

Descrição: Merge das duas tabelas manipuladas, tanto a de "dataset_grade_diaria" (tabela inicial), quanto à "_geral" (concatenação das três tabelas) com base nas features "Data" e "Hora Início". Também usando o método drop, para a remoção de colunas não utilizadas.
Objetivo: Unificar as duas tabelas para conseguir unir a categoria (que está contida na tabela de grade diária) com a audiência (que está contida na tabela da emissora “A” geral).

Limpeza de dados
Célula de Referência: 3.2 Limpeza de dados.

Descrição: Remoção de campos vazios, nulos, ou categorizados como “Não Consta”.
Objetivo: Deixar os dados mais enxutos, tirando campos sem utilidade para o modelo.

Derivação de novos atributos: Mês e Dia a partir da coluna “Data”
Célula de Referência: 3.3 Derivação de novos atributos: Mês e Dia a partir da coluna “Data”.

Descrição: Separação da coluna "Data" para novas colunas "Mês" e "Dia".
Objetivo: Poder associar essas colunas com o mês e dia de um programa que será lançado, permitindo um modelo com mais clareza.

Conversão das Colunas para o tipo “Int”
Célula de Referência: 3.4 Conversão das Colunas para o tipo “Int”.

Descrição: Transformação da Hora Início, em intervalos de 15 minutos, mudando o número de acordo com o intervalo. Transformação da nomenclatura dos dias da semana em números para entendimento do programa.
Objetivo: Poder utilizar as colunas, a partir dos dicionários data e dia da semana, como novos inputs no futuro.

Utilização do Método One-Hot-Encoding para tratar os dados categorizados (Categoria)
Célula de Referência: 3.5 Utilização do Método One-Hot-Encoding para tratar os dados categorizados (Categoria).

Descrição: Cada categoria é adaptada para uma coluna, assim o sistema entende que se naquele instante (dentro do nosso intervalo de 15 minutos) determinada categoria estiver passando, ela será contabilizada com 1 (verdadeiro) e se não estiver, será 0 (falsa).

Objetivo: Associar cada categoria com um número, para o programa interpretar o que está acontecendo.

Agregação da tabela ”Feriados” com a tabela “Emissora A”
Célula de Referência: 4. Agregação da tabela “Feriados” com a tabela “Emissora A”.

Descrição:
Data: Transformação de “string” para “datetime”.
Coluna feriado: Transformamos os valores da coluna feriado em números booleanos.
Objetivo: Trazer os feriados nacionais para incluir como coluna no nosso modelo, tendo assim outra variável, buscando uma melhor predição.

Separação das colunas em features e labels
Célula de Referência: 5.1 Seleção das Features e Labels.

Descrição: Foram criadas duas variáveis, x e y , que armazenam, respectivamente, as features selecionadas para utilização do modelo de regressão, tais como: “Hora Início”, “Dia da Semana”, “Mês”, “Dia”, “Feriado”, Categorias, e outras, além de labels para a tabela final, como por exemplo: “ Rat% Total Domicílios”, Rat% por classe social e Rat% por gêneros.
Objetivo: Ter um núcleo com todas as informações selecionadas, para obter as informações necessárias para a predição de performance de um novo programa.

4.3.3 Agregação de Registros e derivação de novos atributos

O processo de agregação de registro e derivação de novos atributos foi realizado por etapas, por meio da mesclagem de planilhas selecionadas para o modelo preditivo. As planilhas utilizadas foram: “Planilha Grade Diária”, “Planilha Emissora A Geral”, “Planilha Programação” e "Planilha Feriados Nacionais”.

Merge das planilhas

O processo de mesclagem das planilhas foi feito por etapas, começando pela junção das planilhas da Grade Diária e Emissora Geral A, seguindo para a Planilha Programação, e por último a Planilha feriados nacionais. A técnica utilizada foi “left merge”, função pd.merge, na qual se mantém todos os dados da planilha à esquerda, e adiciona-se a nova planilha à direita, transformando suas linhas inexistentes em valores NaN.

Exemplo de código utilizado para mesclagem das tabelas:

```
tabela_programacao = pd.merge(emissora_A_geral, dataset_gradediaria,
how="left", left_on=['Data', 'Hora Início',], right_on =['Data', 'Hora
Início'])
```

Concatenação

Foi-se aplicado o método concat() para juntar dados das planilhas de seg-sex, sábado e domingo. A concatenação agrupou as três planilhas verticalmente, por coluna, de forma não ordenada.

Código utilizado para concatenação dos dataframes:

```
tv_emisssora_A = pd.concat([tv_emissora_A, tv_emissora_A_sab, tv_emissora_A_dom])
```

Derivação de novos atributos

Para o processo de derivação de novos atributos de outras colunas, foi-se aplicado a função dt.strftime(%), a qual cria uma nova coluna a partir de um valor que está no formato datetime. Também foi utilizado a função .split , para separar strings e posteriormente converter para datetime, visando a extração eficiente e formatação dos dados para saída e manipulação.

Exemplo do código utilizado para derivação de novos atributos:

```
tabela_programacao['Mês'] = tabela_programacao['Data'].dt.strftime('%m')
```

One-hot encoding

O método foi aplicado para transformar os valores dos atributos em formato de números binários, aplicando a lógica booleana. A função do pandas utilizada foi pd.get_dummies.

Exemplo do código utilizado no método:

```
Tabela_convertido = pd.get_dummies(tabela_convertido,columns=['Categoria'], drop_first=True)
```

Dicionários

Os dicionários foram criados para agregar intervalos de string em um único número, na coluna de Horas e Dia da semana**.**

Horas: Os intervalos de 15 minutos foram separados em um único número, para identificar aquele horário. Períodos de Xh até Xh10m foram identificados como X; Xh15m até Xh25m foram identificados como X,25; Xh30m até Xh40m foram identificados como X,5; Xh45 até Xh55 foram identificados como X,75.

Dia da semana: Foi numerado de 0 até 6, começando por domingo.

Exemplo do código utilizado:

```
dicionario_dia_semana = {"Domingo":0,"Segunda":1,"Terça":2,"Quarta":3,"Quinta":4,"Sexta":5,"Sábado":6}
```

4.3.4 Remoção e substituição de valores ausentes, em branco, ou desconsiderados

Ao vislumbrar o dataframe, não foram encontradas quaisquer linhas ou colunas com valores nulos ou ausentes, porém, durante a análise, foi verificada a categoria “Não Consta”, considerada como valor ausente, pois não havia nenhuma informação referente a programação. Ademais, como descrito na tabela, foi concluído que essa categoria poderia ser definida como um momento de transmissão ausente, indefinido ou que há apenas a logo da emissora e, por apresentar pouquíssimas quantidades de valores ao comparar com o total da tabela, esse tipo categórico foi considerado irrelevante.

Desse modo, para conhecer e visualizar todos os valores da coluna categoria, foi utilizada a biblioteca Pandas e funções para converter os valores em uma lista, para assim, confirmar se os valores da coluna buscada estavam realmente presentes. Para prosseguir com o processo de remoção, foi efetuado o “drop” e, a partir disso, removida cada linha especificada que obtém o valor “Não Consta”, sem obter uma cópia, no final, filtrando os dados resultantes para ter uma parte da visualização da tabela. Portanto, as funções necessárias para visualizar os valores, assim como o “drop” para realização da remoção das linhas, estão disponíveis nas células textuais do Google Colab.

4.3.5 Identificação das features selecionadas

Como características padronizadas de um novo modelo, de um programa, em que abordam valores de entrada, selecionamos as seguintes features que se encaixam com esses valores:

Hora Início
Utilizado para separar o intervalo de tempo da predição do modelo preditivo, a qual será determinada em intervalos de 15 em 15 minutos, ao invés de a cada 5 minutos, pois assim, há um maior aproveitamento do tempo e mais dados são encapsulados devido a duração estendida.

Dia da Semana, Mês e Dia
As três features, em conjunto, serão utilizadas para identificar os respectivos “dia da semana”, “mês” e “dia” para qual a audiência da predição será calculada. Estas foram formatadas e divididas em colunas diferentes, para possibilitar a leitura dos valores pelo sistema.

Feriado
Os valores das colunas foram transformados em números booleanos, para a leitura do sistema, e integrados com a tabela original. Essa integração foi considerada, pois ocorrem mudanças na grade da programação, e os feriados possuem influência positiva na audiência, apresentando-se relevantes.

Categoria

Anteriormente, as categorias estavam em formato string e em linhas. Foi executado processos de formatação e conversão das linhas categóricas, em colunas, assim como a remoção da primeira linha da tabela chamada “categoria”, não mais necessária. Os valores que fazem parte dessa categorização, foram convertidos em números de identificação, para possibilitar a leitura pelo sistema. Por exemplo:

- Categoria 1: valor 1;
- Categoria 2: valor 2;
- Categoria 3: valor 3.

Desse modo, totalizou-se 46 colunas, utilizando os tipos de categorização específicas para cada programa na grade.

Labels

Para obter a audiência dos programas, é efetuada a medição pelo Rat%, e os labels são os resultados que devem ser devolvidos, servindo como um objeto extra a ser inserido.

Não são features, mas são outputs/targets esperados para o modelo desenvolvido. Os labels utilizados são:

- Total Domicílios;
- Rat%, AB;
- Rat%,C1;
- Rat%, C2;
- Rat%, DE;
- Rat%, Masculino;
- Rat%, Feminino;
- Rat%, 4-11 anos;
- Rat%, 12-17 anos;
- Rat%, 18-24 anos;
- Rat%, 25-34 anos;
- Rat%, 35-49 anos;
- Rat%, 50-59 anos;
- Rat%, 60+ anos | Rat%.

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

CUI, Jingsong; SEREDAY, Scott; , VP. Using Machine Learning to Predict Future TV Ratings In An Evolving Media Landscape. Nielsen, 2016

MA, Nan. Prediction of Television Audience Rating Based on Fuzzy Cognitive Maps with Forward Stepwise Regression: International Journal of Pattern Recognition and Artificial Intelligence, 2016.

BONDADE, Navi. With AI Fox Film Studio Predicts Movie's Audience By Analyzing It's Trailer. India, 2018.

Lightreading. Gracenote launches 'Audience Predict' tool to gauge content performance. Emeryville, Calif, 2021.

LIAO, Shannon. BBC will use machine learning to cater to what audiences want to watch. New York, 2016.

VAZÉ, Achyut. Building a Model for Predicting TV Ratings. Mumbai, 2016.

XIA, Jhiazi. TVseer: A visual analytics system for television ratings. Hangzhou, 2020.

AKULA, Ramiya, WIESELTHIER, Zachary, MARTIN, Laura, GARIBAY, Ivan. Forecasting the Success of Television Series using Machine Learning, Orlando, USA, 2019.

Istoé Dinheiro. BRASIL empobrece em 10 anos e tem mais da metade dos domicílios nas classes D e E, 2022.

Anexos

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