PredTv
Rede Gazeta

inteli
instituto
de tecnologia
e liderança

Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
08/08/2022	Antonio Teixeira Sophia Tosar	1.1	Adição do Canvas de Proposta de Valor e Análise SWOT.
10/08/2022	Giovanna Rodrigues	1.2	Adição da Matriz de Risco.
11/08/2022	Giovanna Rodrigues	1.3	Adição das seções 1 e 2.
12/08/2022	Giovanna Rodrigues Sophia Tosar	1.4	Adição da seção 4.1
26/08/2022	Antonio Teixeira Daniel Barzilai Giovanna Rodrigues Sophia Tosar	1.5	Adição seção 4.2 e 4.3
08/09/2022	Antonio Teixeira Daniel Barzilai Sophia Tosar	1.6	Adição seção 4.4

Sumário

1. Introdução

A Rede Gazeta, fundada em 1928 com a primeira edição impressa do jornal A Gazeta, atualmente, é composta pelo site de notícias A Gazeta, pelas rádios CBN Vitória, Gazeta FM, Rede Litoral e Mix Vitória, pelas quatro emissoras da TV Gazeta (Grande Vitória, Norte, Noroeste e Sul) e pelos portais g1 ES e GE ES, e o maior grupo de comunicação multimídia do Espírito Santo. Afiliada à TV Globo, a emissora TV Gazeta atinge 41% da população capixaba e tem como objetivo contribuir com o desenvolvimento e fortalecimento de seu estado.

Acerca de sua programação, por ser uma afiliada, ela deve seguir a grade principal da TV Globo, mas tendo slots disponíveis para os programas de sua preferência. Nesse contexto, mesmo possuindo dados históricos de audiência, a TV Gazeta ainda não possui uma forma de prever se a aceitação da população aos novos programas esperada é superada, causando ansiedade e insegurança aos responsáveis pela programação.

2. Objetivos e Justificativa

2.1. Objetivos

O objetivo da TV Gazeta principal é o retorno dos seus investimentos em novos programas, logo, para que isso aconteça, é imprescindível que o novo programa tenha uma boa audiência e um bom score.

Nesse objetivo, eles devem entender quais características, como horário e gênero, mais influenciam na notoriedade de seus novos produtos, para assim, conseguir criar novos programas que vão a favor de todas essas características, tendo maior certeza de sucesso.

2.2. Justificativa

Neste projeto será descrito o PredTv, uma plataforma baseada em *machine learning capaz* de prever pontuações de audiência de programas futuros de acordo com as características escolhidas pela produtora, além de conseguir identificar quais dessas características têm mais peso sobre o score fornecido. Assim, a produtora consegue ter mais claridade sobre os elos fracos de seu novo produto e fortalecê-los com antecedência.

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

Principais players

O público-alvo deste projeto é o setor de inovação da Rede Gazeta, que utilizará o modelo preditivo criado para antever a audiência de produtos criados de acordo com seu horário de exibição, data, idade, gênero e classe social do público-alvo. Uma vez que a previsão será direcionada à audiência da Rede Gazeta, serão considerados para análise os telespectadores da TV Gazeta. Tal público pode escolher livremente assistir qualquer canal de TV disponível em seu domicílio. Portanto, para esta análise de mercado, foram consideradas as três maiores emissoras de TV aberta além da Rede Globo, visto que A Gazeta é sua afiliada.

De acordo com o levantamento TV PNT TOP publicado pela Kantar IBOPE Media, as maiores emissoras de TV do Brasil em audiência no ano de 2021 foram: Globo: 10,11, Record: 4,27, SBT: 3,39 e TV Band: 0,96. A receita líquida destas mesmas emissoras no ano de 2021 foi de: Globo Comunicação e Participações S.A: R$14,4 bi, Sistema Brasileiro de Televisão: 1,4 bi, Record TV Rádio e Televisão Record S/A: R$ 130,503 mi (lucro líquido).

Modelo de Negócios

Merchandising – Mechan é o momento em que o apresentador para o programa para apresentar o produto. Assim, esta apresentação é paga pela empresa que publica o produto. Scripts com as marcas dos patrocinadores e textos sobre o produto fazem com que o merchandising também gera, naturalmente, renda para a emissora.

Espaço Publicitário – Os espaços publicitários são momentos de pausa na programação para os horários comerciais. Esses comerciais existem desde 1951, eles entraram um ano após a chegada da televisão no Brasil.

Aluguel de Horário – Aluguel de horário são momentos em que programas de televendas ou programas de igrejas alugam um tempo da programação. Dessa forma, alugar um horário é uma forma de obter uma receita fixa para a emissora.

Tendências

Mesmo com o crescimento do uso de serviços de *streaming*, 79% da população brasileira absorve conteúdos audiovisuais por meio de TV aberta e canais por assinatura.

5 Forças de Porter

Ameaça de produtos substitutos:

Atualmente, vivemos na era da tecnologia, logo, serviços de streaming se tornaram cada vez mais populares, pois oferecem uma variedade de conteúdos e são mais práticos, uma vez que os usuários têm acesso por outros dispositivos, como computador e celular. Contudo, não se espera que isso se torne uma grande ameaça, com a criação da sua própria plataforma de streaming, a Globoplay, a Globo se inseriu nesse mundo da tecnologia e têm sucedido. No último trimestre deste ano, a arrecadação financeira pela Globoplay aumentou em 50%, animando a rede televisiva. Além do fato das televisões persistirem em quase todas as casas brasileiras e terem seu consumo impulsionado com a pandemia, portanto, se torna improvável a migração de usuários para outros produtos além do espectro da Rede Globo.

Ameaça de entrada de novos concorrentes:

Entrar no ramo de televisão pode ser muito intenso, isso se deve aos altos custos de investimento, mas também, deve-se considerar a força das marcas já estabelecidas, uma vez que estão nesse mercado por muitos anos e já possuem o respeito e fidelidade da população, possuindo quase um monopólio na tv aberta.

Poder de negociação dos clientes:

Devido a grande polarização política atual, se observa uma grande movimentação contra as emissoras contrárias à ideologia de seus usuários e com isso, eles migram para emissoras concorrentes. Se observa que nos primeiros quatro meses deste ano, o Jornal Nacional possuiu o menor número de audiência de sua história, jornal que é frequentemente afrontado por questões ideológicas.

Poder de negociação dos fornecedores:

Os fornecedores dessa indústria são aqueles que produzem os programas televisionados. Nesse âmbito, possuímos duas categorias, os fornecedores nacionais e internacionais, entre esses, o poder de negociação é contrário. Entre os fornecedores nacionais, esse poder é pequeno, uma vez que a possibilidade de terem seus produtos transmitidos por emissoras de grande porte é diminuta. Enquanto os fornecedores internacionais, possuem a chance de distribuir seus serviços para um maior número de consumidores, como plataformas de streaming e diversas emissoras, possuindo um poder maior.

Rivalidade entre os concorrentes:

As grandes concorrentes estão estabilizadas no mercado por muitos anos, possuindo um monopólio de audiência, contudo, nota-se que mesmo com a migração de espectadores de uma emissora para outra, a Rede Globo continua possuindo o primeiro lugar entre elas. Um exemplo, é o fato mencionado anteriormente, apesar da grande perda de audiência pelo programa Jornal Nacional, ele continua sendo o maior telejornal brasileiro. A rivalidade é grande, já que o câmbio entre artistas é dificultado ou barrado, mas é inegável que o poder da emissora Globo é superior ao das outras emissoras.

4.1.2. Análise SWOT

Ambiente Interno	**FORÇA**	**FRAQUEZA**
	1.A criação do modelo preditivo pode auxiliar na medição de audiência. 2. A Rede Gazeta atende 41% da população capixaba. 3. Grande grupo de comunicação e multimídia do Espírito Santo.	1.O People Meter pode gerar dados enviesados. 2. Possuem um grupo pequeno na área de inovação 3. Prioriza o estado do Espírito Santo, que é um estado pouco populoso
Ambiente Externo	**OPORTUNIDADE**	**AMEAÇA**
	1. Aumento da audiência da Rede Gazeta com a criação do modelo preditivo. 2. Fortalecimento do desenvolvimento do Espírito Santo.	1. Competitividade de mercado 2. Aumento do uso de smartphones que afeta o tempo dos telespectadores conectados na TV.

4.1.3. Planejamento Geral da Solução

Ao lançar novos conteúdos, a Rede Gazeta deve decidir diversas variáveis que influenciam na audiência de seu produto, apesar de possuírem uma quantidade elevada de dados históricos de audiência, não é possível ter extrema certeza sobre um score futuro, logo, há muita insegurança no investimento de lançamentos, uma vez que há o medo de que não haja retorno financeiro.

Dentre esses dados históricos, tivemos acesso aos últimos três anos de audiência da própria TV Gazeta divididas em métricas importantes na área televisiva (mais contextualizadas na área 4.3 do documento) e entre diversos perfis sociais, como idade, gênero e classe socioeconômica. Assim como a de outras 4 concorrentes para comparação. Todos os dados foram fornecidos pela Rede Gazeta e devem permanecer confidenciais.

A partir disso, desenvolvemos o PredTv, um modelo preditivo que antecipa a audiência de acordo com variáveis adicionadas pelos usuários. Nosso produto vai ajudá-los a saber se um programa, que ainda não foi lançado, vai atender as expectativas esperadas. Além de informar o peso que cada variável tem sobre o score final, assim, a TV Gazeta tem maior controle sobre quais características do novo produto devem ser modificadas para uma maior aceitação.

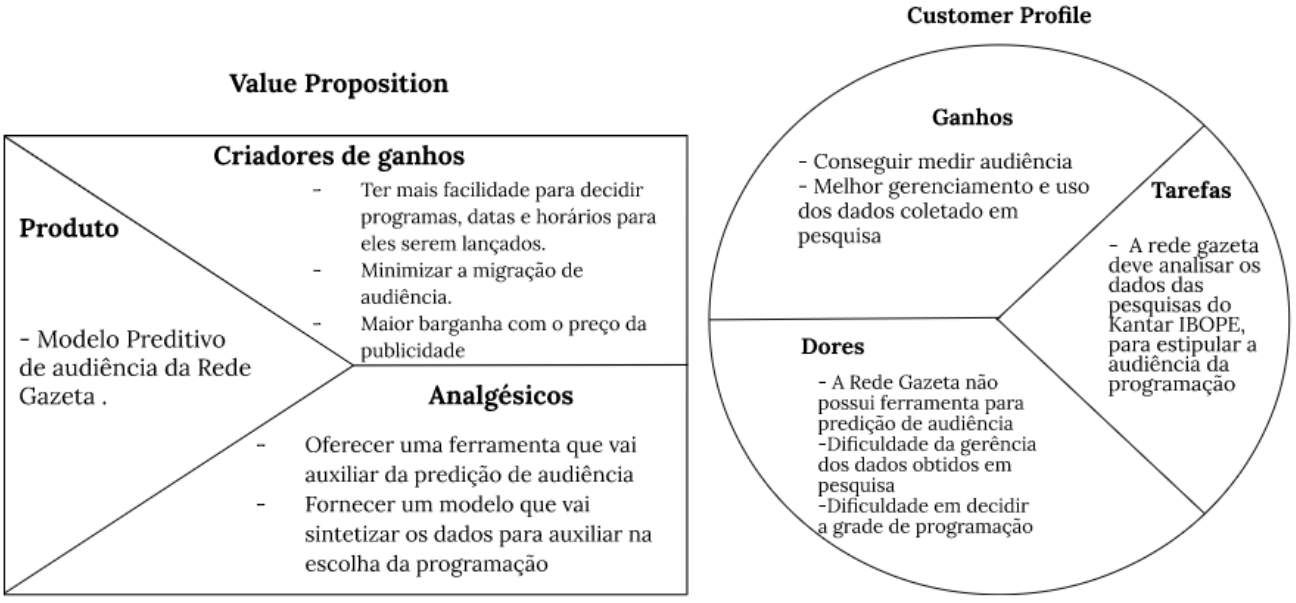
O tipo de tarefa utilizado será regressão, pois nós devemos prever uma resposta, estimando um valor numérico.

Posteriormente, o PredTv deverá ser utilizado pela equipe de marketing e programação da Rede Gazeta, de maneira que eles irão imputar as variáveis desejadas (horário, data, público-alvo) e o programa gerará um score baseado nesses inputs. Ao receber o resultado, a equipe deve ponderar se ele é o esperado e, se não, mudar as variáveis que mais influenciaram, elas serão informadas pelo próprio programa, até que o resultado seja satisfatório.

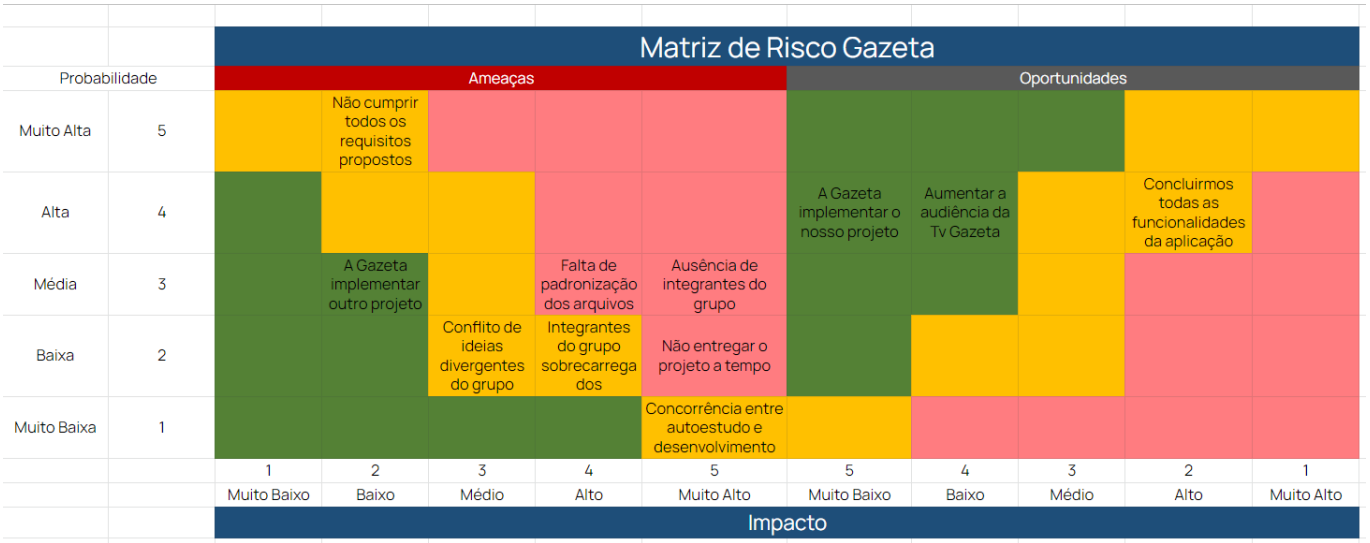
Consequentemente, a emissora terá maior facilidade na decisão das variáveis de um novo programa; terá um melhor gerenciamento e uso dos vários anos de pesquisa de audiência, podendo até diminuir esse armazenamento no futuro, uma vez que o programa já saberá o impacto das variáveis no seu resultado a partir dos anos usados anteriormente; conseguirá diminuir a migração de telespectadores; além de ter maior poder de barganha nos preços de publicidade e maior retorno financeiro.

Ao final, a solução será considerada um sucesso quando sua margem de erro for pequena ou desconsiderável.

4.1.4. Value Proposition Canvas



4.1.5. Matriz de Riscos



4.1.6. Personas

Luis Cassius Gomes, 36 anos, Diretor de inovações.

Biografia: Formado em tecnologia na área de análise e desenvolvimento de dados; fez pós-graduação em gestão corporativa; trabalha numa rede de TV; está há mais de 12 anos no ramo.

Características: Costuma buscar soluções através de tecnologia; gosta de estar envolvido com processos de inovação; busca soluções em automatização de processos; procura sempre manter-se atualizado; tem boa conexão com os coordenadores e diretores de TV.

Motivações com a plataforma: Melhorar a audiência dos programas da empresa; melhorar a assertividade na escolha da programação.

Motivação com o problema: Não possui um sistema customizado e flexível de predição para os novos produtos de sua programação..

Dores: Possibilidade de não estar aproveitando 100% o potencial de um programa; não possui uma aplicação auxiliar ao PeopleMeter; falta de métricas mais acuradas; ausência de um sistema no qual ele possa manipular os dados à vontade.

Santos da Silva, 60 anos, Apresentador e Produtor-chefe.

Biografia: Formado em Produção audiovisual; ganhou diversos concursos de curta-metragens antes de se formar; começou por baixo na empresa, sendo promovido até seu cargo atual; apresenta um programa na rede de maior audiência do estado.

Características: Apaixonado por arte audiovisual; personalidade cômica; domina a arte da liderança; possui entendimento amplo sobre a produção e desenvolvimento de programas de comédia.

Motivações com a plataforma: Saber quais programas priorizar (os que possuem maior audiência); aumentar a audiência de seu próprio programa.

Motivação com o problema: Não sabe o que deve ser feito para melhorar a audiência.

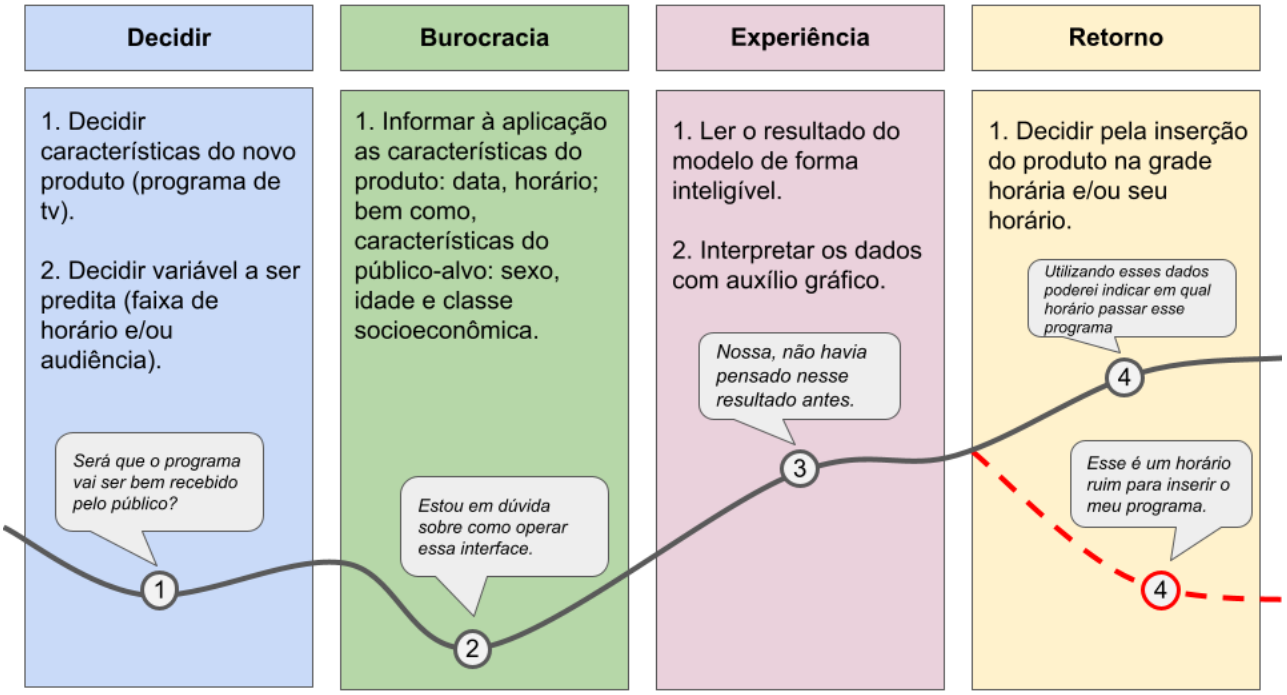
Dores: Não possui um sistema para medir o melhor programa com base nas audiências e público-alvo; carência de um modelo com informações de fácil leitura e compreensão.

4.1.7. Jornadas do Usuário

Luis Cassius Gomes

Cenário: Diretor de inovações testa a plataforma de predição para a predição do *score* de audiência de um novo programa sendo considerado para integrar a grade horária.

Expectativas: Diretor de inovações recebe, sem dificuldades, como resposta, a predição de audiência de acordo com as características informadas. No mesmo resultado também está presente o nível de influência que cada tipo de dado teve no resultado final.



O mapeamento demonstrado acima indica que temos como oportunidade a criação de uma interface para facilitar o uso do modelo criado. Desta forma, a experiência do usuário durante a realização das tarefas será mais intuitiva e agradável. No que diz respeito ao modelo, um efeito secundário da possibilidade de prever a audiência de forma qualitativa em faixas de horário específicas é a escolha de horários de publicidade. Por fim, uma vez que a plataforma for implementada em seu cenário ideal de utilização, esta poderá avaliar se suas predições estão de acordo com o comportamento da audiência no domínio empírico. Sendo assim, identifica-se a oportunidade de usar o confronto das predições com os dados empíricos recém-coletados para retroalimentar o modelo, tornando-o mais acurado.

4.2. Compreensão dos Dados

Dados fornecidos pela Rede Gazeta sobre seu histórico de audiência e grade de programação e a de seus concorrentes principais. Os dados informam valores de métricas importantes no mundo da televisão, tais como:

- rat (televisores conectados nessa emissora dentre os televisores possíveis)
- share (televisores conectados nessa emissora dentre os televisores ligados)
- fid (televisores que permaneceram conectados nessa emissora por mais de um minuto).

Eles foram medidos de 5 em 5 minutos por dois anos e categorizados em classes econômicas (AB, C1, C2, C3, DE), gêneros (feminino e masculino) e idades. Foram 438 MB de dados em formato XLSX.

A partir disso, podemos agregar os horários de audiência com a programação, dados que também foram enviados pelo cliente em uma nova planilha de formato XLSX e 5 MB de tamanho , que possui o nome dos programas juntos de sua categorização, monitorados de 5 em 5 minutos no período de 2 anos, e assim definir a relação entre telespectadores, horário e tipo de conteúdo.

A pesquisa é feita pela empresa Kantar IBOPE Media, de renome na área de pesquisas de audiência. Os dados são coletados através de dispositivos chamados PeopleMeter, que coletam o perfil do espectador e a emissora assistida. São, em média, 200 aparelhos distribuídos por residências com diversos perfis e que rotacionam a cada 2 anos para evitar algum viés na projeção.

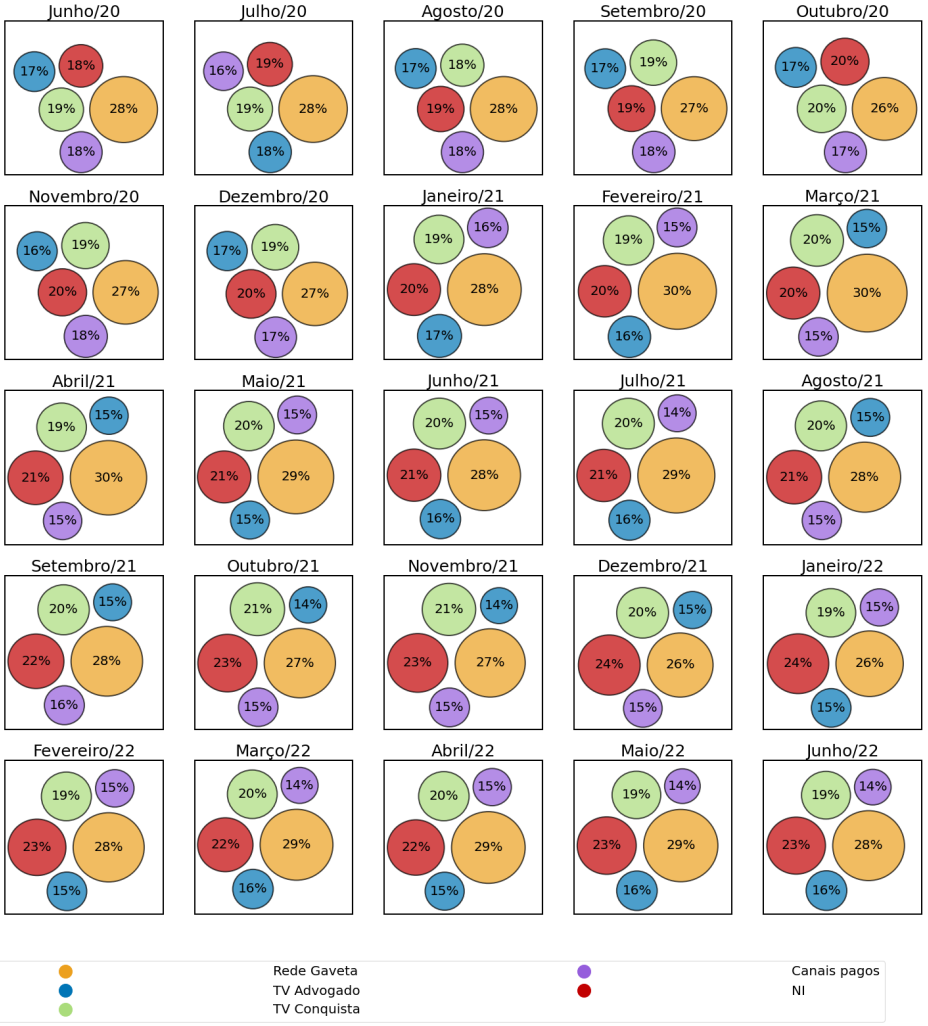
Posteriormente, fizemos subconjuntos para uma análise mais profunda, sendo esse:

- Planilha composta apenas pelo rat voltado para domicílios de todas as emissoras, a métrica de maior importância na análise de audiência. Podemos então, determinar horários de pico, médias de audiência de um nível geral.
- Planilha composta apenas pelo rat voltado para os gêneros, podendo então, determinar as médias de audiência por gênero e seus horários de pico.

Os dados são confidenciais, logo, não serão expostas os nomes das emissoras nesse documento.

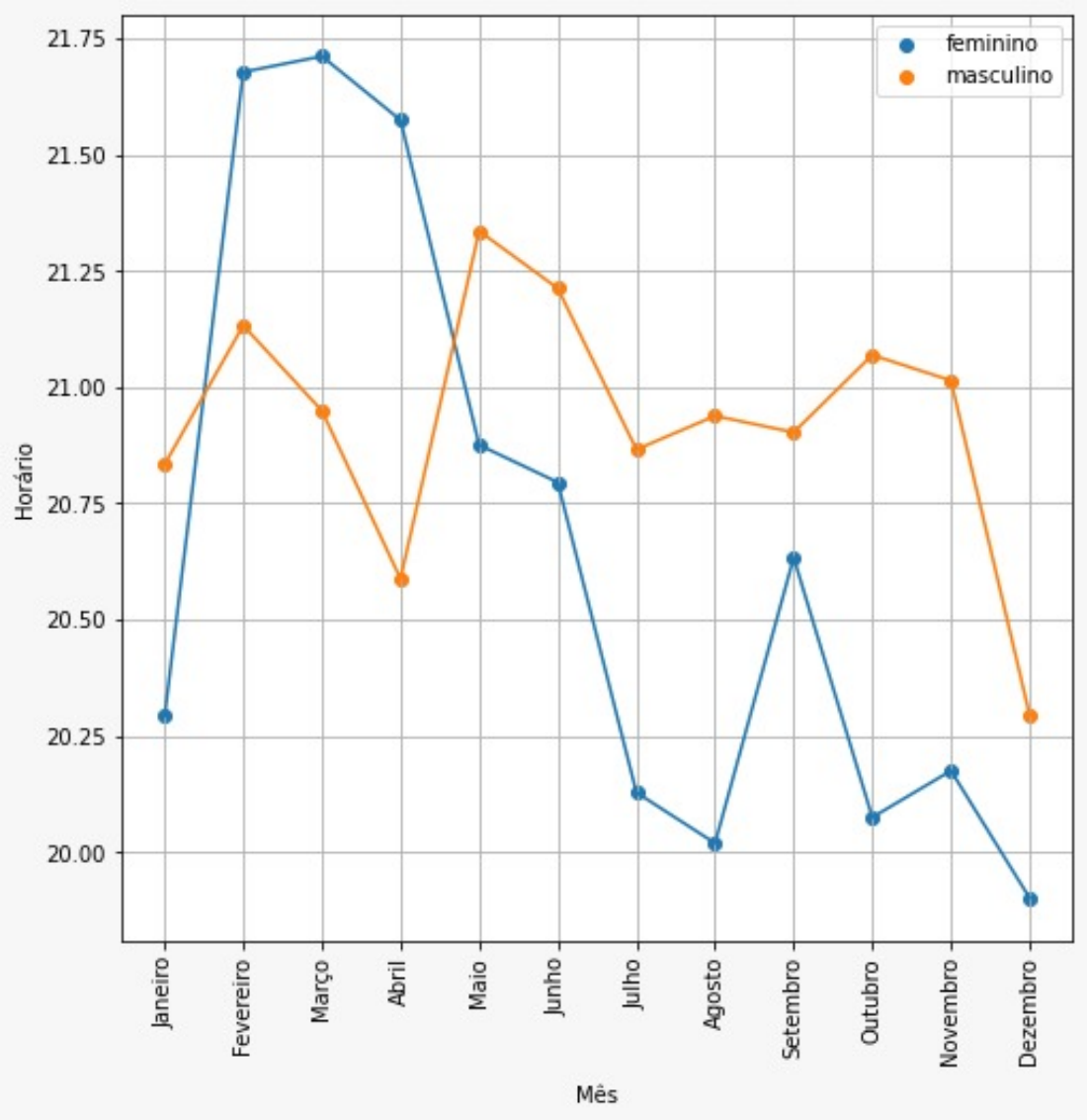
O resultado (*target*) dessa predição será o score de audiência, por ser um número aproximado da realidade e não um resultado binário entre “audiência boa” ou “audiência ruim”, sua natureza é contínua.

Gráfico de audiência das emissoras do Espírito Santo



O gráfico acima nos informa a média mensal nos anos de 2020 a 2022, é possível a partir daí, determinar que a Rede Gaveta possui uma predominância na maior audiência, sendo, em todos os meses, a maior audiência. Ademais, também é possível afirmar que as emissoras NI e Rede Conquista disputavam pela segunda colocação por muitos meses, com audiências iguais ou com diferença de um porcento, contudo, a partir do mês de setembro de 2021, a emissora NI possuiu uma diferença de dois porcento que vem aumentando desde então, demonstrando um fator de tendência nos últimos meses.

Gráfico de diferença de audiência por gênero



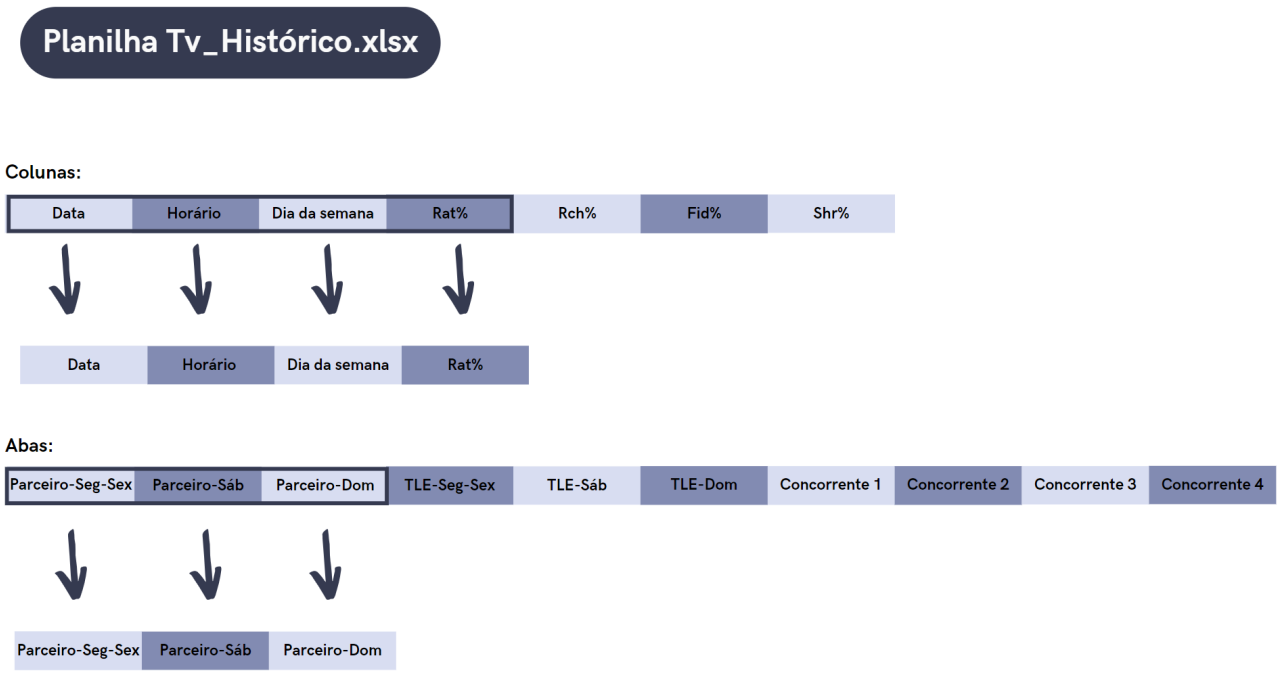
No gráfico acima, os dados demonstrados são resultados do horário com a maior média de audiência de todos os meses de janeiro, fevereiro e assim por diante. Inicialmente, achamos interessante calcular a média de audiência por horário em meses iguais, uma vez que supomos que esses possuiriam programações similares, contudo, esse gráfico está enviesado, já que meses iguais podem possuir programas cíclicos, logo, um programa passado no primeiro mês de janeiro desses dados, não estará mais passando no próximo mês de janeiro e será voltado para um público-alvo distinto, então, os horários de pico entre meses iguais são discrepantes e uma média entre eles não traria valores reais. Portanto, um novo gráfico será feito com as médias de horário de pico de todos os meses, trazendo mais sustentação para nossas hipóteses.

4.3. Preparação dos Dados

4.3.1. Otimização dos dados

Os dados iniciais possuíam um conteúdo muito grande, logo, o seu carregamento demandava muito tempo. Então, para a preparação dos dados foi necessário, inicialmente, a realização de uma nova planilha, em formato CSV, para que a manipulação dos dados disponibilizados pelo parceiro fosse feita de forma mais dinâmica e com o objetivo de obter um artefato que contivesse apenas os dados que iremos, efetivamente, utilizar na criação do modelo preditivo, tendo seu processamento agilizado.

O novo arquivo gerado foi feito com base em três abas, pertinentes ao parceiro, da planilha original "TV_Histórico.xlsx", uma com todos os dados referentes apenas aos dias úteis, enquanto as outras duas eram referentes ao sábado e domingo. Logo após, as colunas selecionadas para o novo arquivo foram aquelas referentes ao "rat%", métrica preferencialmente utilizada na medição de audiência. Consequentemente, as colunas de "shr%", "fid%" e "rch%" foram descartadas. Assim, é obtido um arquivo contendo apenas as informações essenciais em um único lugar.

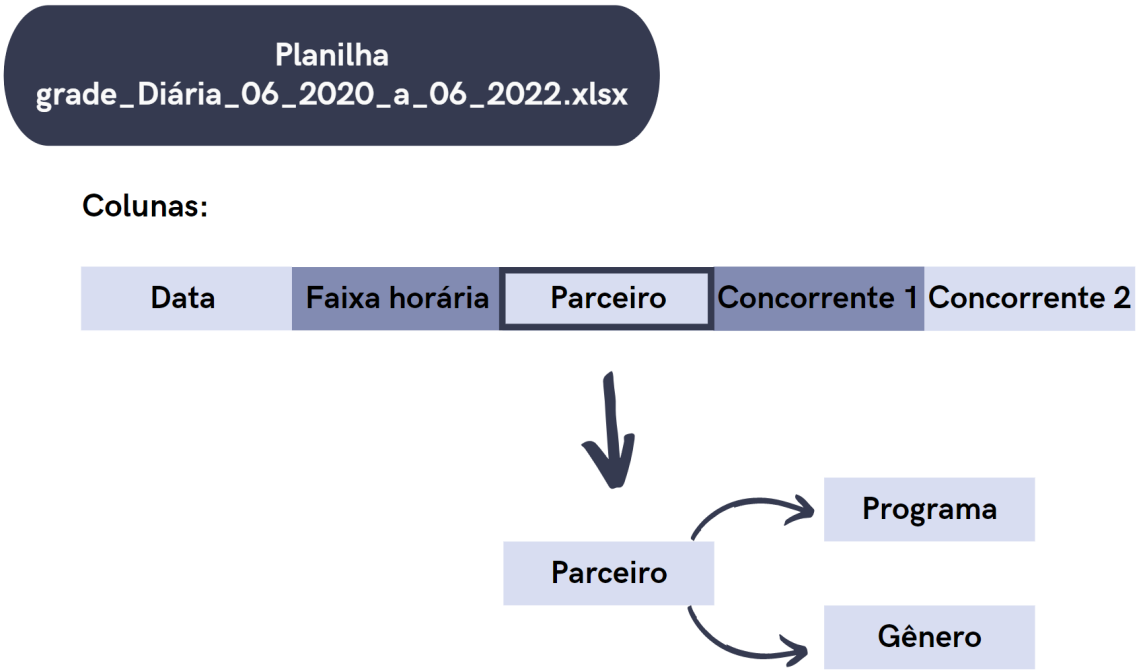


4.3.2. Ordenação das datas

Ao adicionar as abas de final de semana no novo CSV, todos os sábados e domingos são posicionados no final da planilha, dificultando a visualização, uma vez que as semanas não possuem uma linearidade. Então, foi aplicada uma ordenação nas datas, movendo os sábados e domingos para suas devidas posições de acordo com o dia, mês e ano.

4.3.3. Junção da grade horária

Até o momento, o arquivo principal continha apenas valores de audiência e espaços de tempo, mas não possuía a programação exibida. Então, através de outro documento xlsx “grade_Diária_06_2020_a_06_2022” fornecido pelo parceiro, foram movidos os dados a respeito do nome da programação e sua categoria para o novo CSV. Contudo, essas duas informações estavam numa única coluna, dificultando a visualização, então, foi necessário fazer a divisão destas em duas colunas: “Programa” e “Gênero”.



4.3.4. Verificação de valores nulos ou ausentes

Foi feita uma verificação por valores nulos, mas, nenhum foi encontrado, logo, não foi necessário ser feito algum procedimento para removê-los ou substituí-los.

4.3.5. Seleção de Features

Para atingir o objetivo do PredTv, que é prever audiências de novos programas, é importante a análise do histórico, assim, é possível definir as features, características que influenciam diretamente no score final. Nas que se referem a espaço de tempo, as mais importantes seriam data, dia da semana e horário de início do programa, uma vez que essas determinam a quantidade de pessoas possíveis de serem telespectadores naquele momento. Já as voltadas para a medição de audiência, temos a porcentagem do número de domicílios conectados na emissora estudada, o rat, mas também, a mesma métrica para perfis específicos de gênero, idade e classe social, sendo útil para o estudo do impacto do horário e programação em cada um desses perfis. No que tange à programação, sua categorização é

essencial para a realização da relação entre essa e os vários tipos de perfis, buscando entender quais perfis são predominantes em cada categoria.

4.4. Modelagem

Para a geração de modelos foram feitas algumas alterações nas planilhas e nos dados do CSV principal. Os outliers foram corrigidos criando mais duas colunas no nosso CSV principal, para realocar os outliers principais que seriam: audiência do reality show da emissora parceira e a audiência nos dias de feriados nacionais.

Ademais, para facilitar a comparação com os dados das outras emissoras foi realizado o processo de normalização para alguns modelos, que consiste em limitar os dados em um certo alcance gerando melhor proporcionalidade entre os dados das emissoras. Por fim, o único processo que foi realizado para a criação de todos os modelos utilizados foi a geração do CSV definitivo o que é descrito na seção 4.3 Separação de dados.

4.4.1 LGBM

Light GBM é um framework de algoritmo. Gradient Boosting é um método de Machine Learning para complicações de regressão e classificação, que apresenta um modelo preditivo, configurado em um conjunto de modelos de previsão fracos, principalmente as árvores de decisão. Como outros métodos de reforço, ele produz o modelo em partes, e os generaliza, possibilitando o aprimoramento de uma função de perda diferenciável arbitrária.

O principal objetivo do Light GBM é gerar uma cadeia de modelos fracos, no qual cada modelo tem como finalidade, reduzir ao máximo o erro do modelo prévio, mediante uma função de perda. Nos reparos de cada modelo fraco é multiplicado um valor, a taxa de aprendizagem. Tal valor é responsável por instituir o impacto de cada árvore no modelo final. Quanto menor o valor, menor a colaboração de cada árvore.

4.4.2. KNN

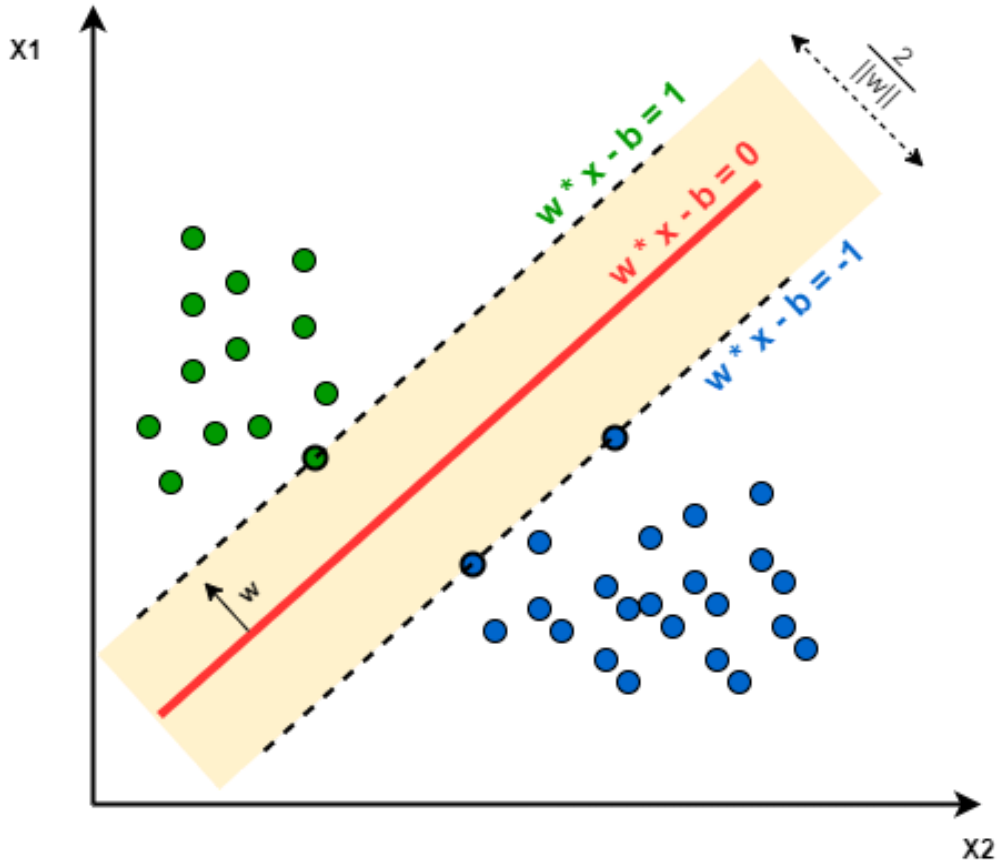
KNN, ou K-vizinhos mais próximos, consiste em um modelo de predição supervisionado baseado na distância com os seus vizinhos mais próximos. Primeiramente, um certo dado é escolhido, e com base em seus atributos, o KNN faz uma comparação com seus vizinhos mais próximos para definir o valor previsto. O k em questão representa o número de vizinhos que serão comparados com o dado escolhido que possui muita influência na comparação, tornando-a mais sensível quando baixo e menos sensível quando mais alto.

As vantagens que contribuíram na escolha do KNN como um dos modelos testados são: a simplicidade de implementação e a sua eficácia em diversas situações.

4.4.3. Support Vector Machine

Support Vector Machine é um algoritmo de aprendizado de máquina supervisionado que pode ser usado para desafios de classificação ou regressão. O objetivo desse algoritmo é criar a melhor linha ou limite de decisão que possa segregar o espaço n-dimensional em classes para que possamos facilmente colocar o novo ponto de dados na categoria correta no futuro. Esse limite de melhor decisão é chamado de hiperplano.

O SVM escolhe os pontos/vetores extremos que ajudam na criação do hiperplano. Esses casos são chamados de vetores de suporte. A imagem a seguir ilustra a forma como o SVM funciona:

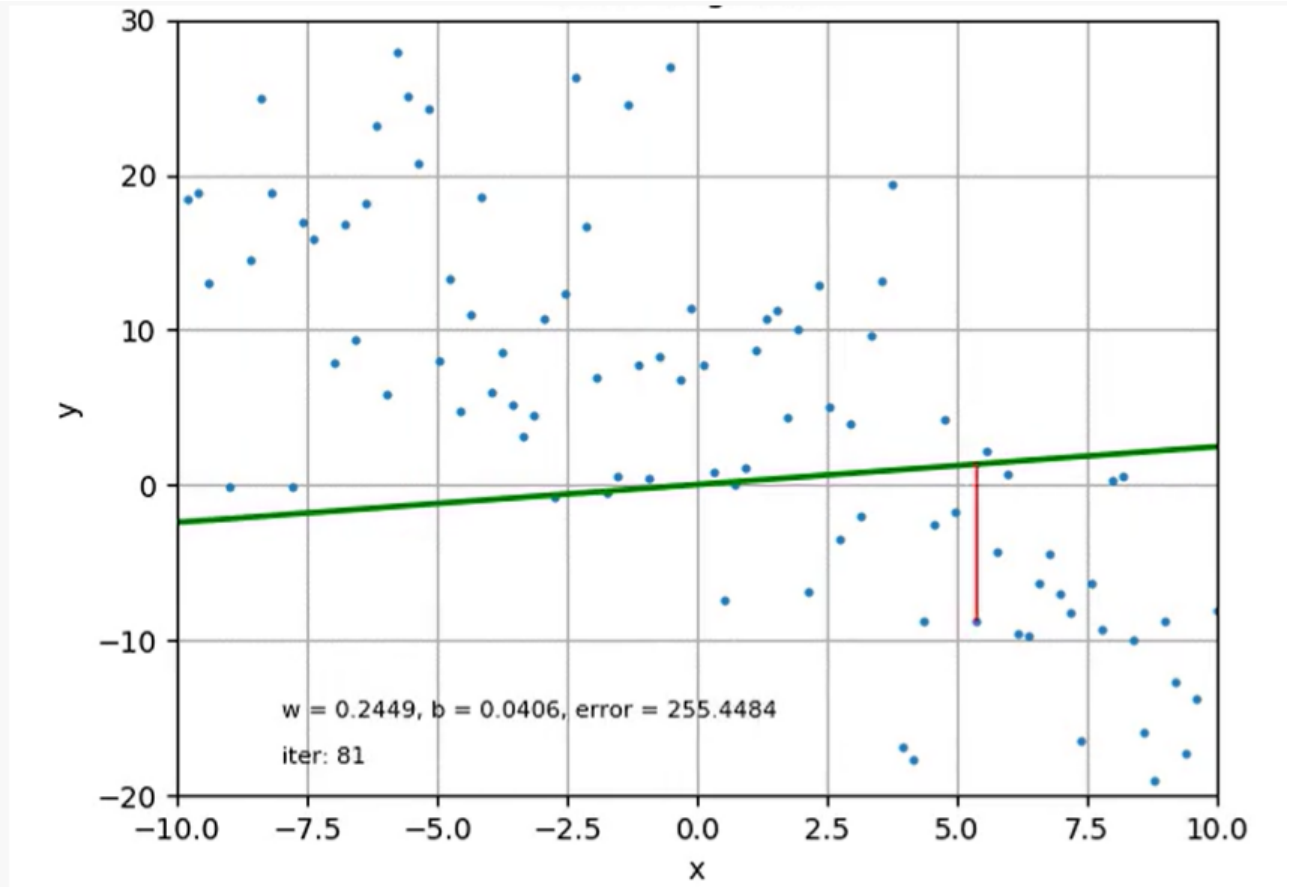


É importante ressaltar que o SVM foi utilizado, pois em caso de outlier, esse modelo busca a melhor forma possível de classificação e, se necessário, desconsidera o outlier. Ademais, o SVM funciona bem em domínios complicados, em que existe uma clara margem de separação.

4.4.4. Linear Regression

Regressão linear consiste em um processo de traçar uma reta através dos dados disponibilizados em um diagrama de dispersão. A reta, quando bem acurada, pode gerar um resumo dos dados, permitindo a criação de um modelo supervisionado preditivo de dados. Esse tipo de modelo foi escolhido por se tratar de um algoritmo de predição mais usual e mais simples de ser feito, ele é bem útil quando não se possui muitos outliers e isso ocorre porque quando existem muitos pontos fora da curva, a reta fica refém dos pontos mais afastados.

Os dados foram submetidos ao processo de normalização padrão, que consiste em limitar os dados entre 0 e 1, e esse processo é útil, pois deixa os dados dimensionados da mesma forma o que resulta em uma melhor relação de proporcionalidade entre os valores das diferentes emissoras.



4.4.5. Random Forest

O Random Forest é um algoritmo que cria muitas árvores de decisão, de maneira aleatória. Cada árvore será utilizada na escolha do resultado final. As árvores de decisão estabelecem regras para a tomada de decisão. Dessa forma, o modelo criará uma estrutura

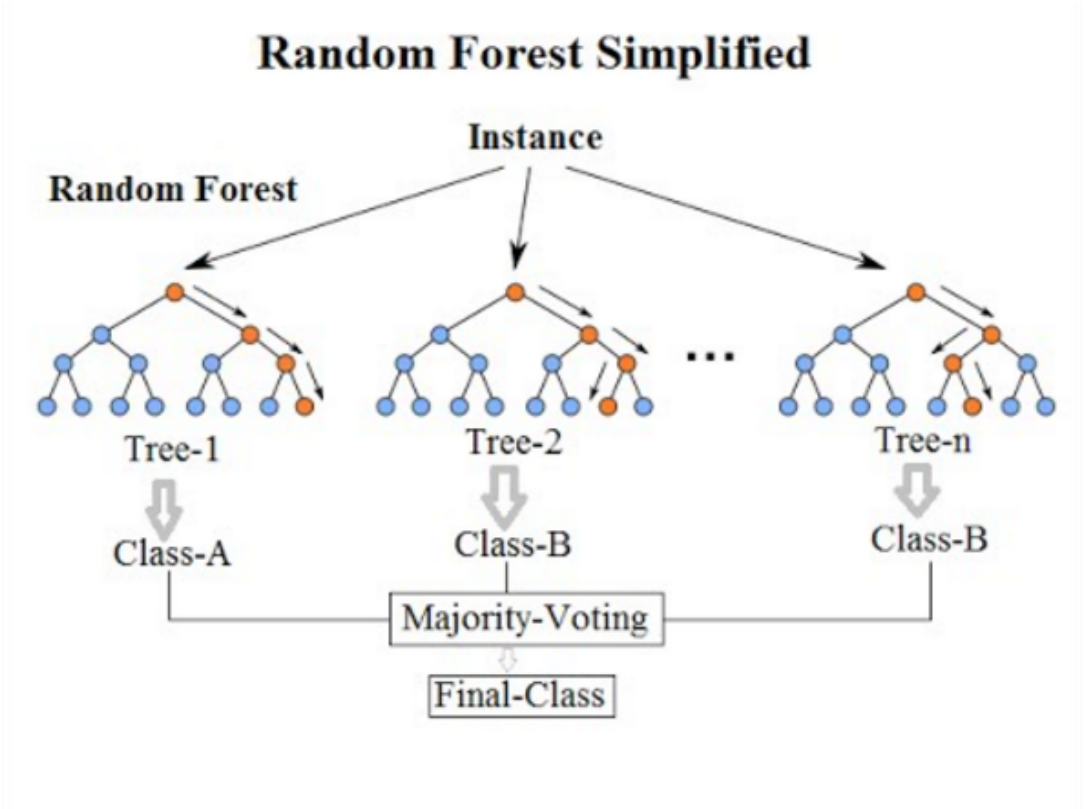
similar a um fluxograma, com “nós” onde uma condição é verificada, e se atendida o fluxo segue por um ramo, caso contrário, segue por outro.

Esse modelo possui quatro etapas principais:

1. Seleção aleatória de algumas features
2. Seleção da feature mais adequada para a posição do nó raiz
3. Geração dos nós filhos
4. Repetir os passos acima até atingir a quantidade de árvores necessárias

Depois que o modelo é criado, as previsões são feitas a partir de “votações”, cada árvore toma uma decisão com base nos dados apresentados, logo a decisão mais votada é a resposta do algoritmo.

O Random Forest foi utilizado, pois tem suas origens na forma mais básica e inicial de um algoritmo de suporte à decisão. Além disso, esse modelo resolve problemas de regressão e de classificação e costuma apresentar bons resultados. A seguir é mostrado, por meio de um esquema, como funciona o Random Forest:



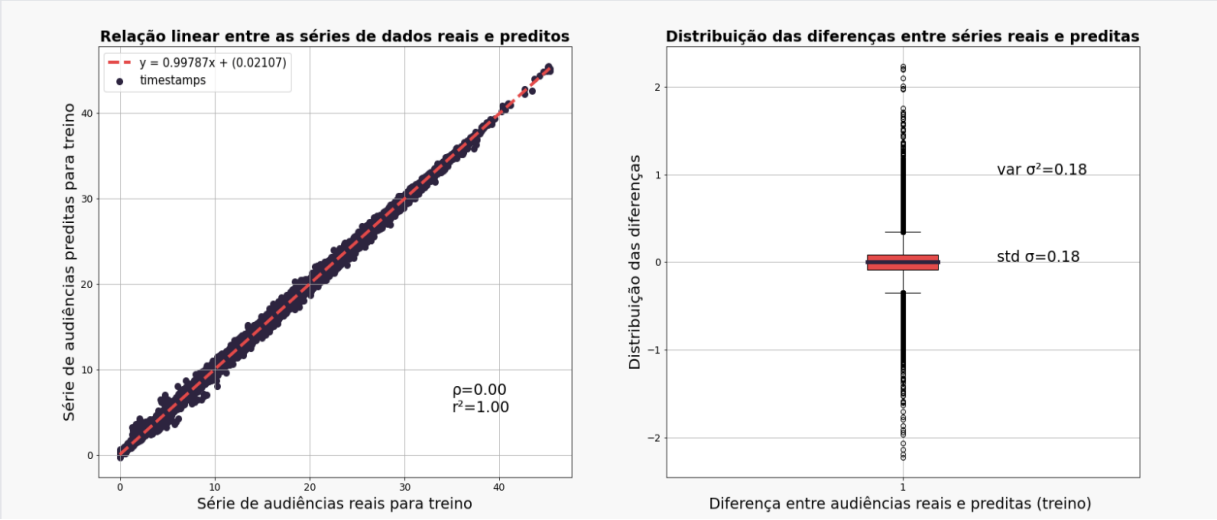
4.5. Avaliação

As variáveis utilizadas para estudar os modelos de regressão foram: Coeficiente de determinação (r^2), que é uma medida de ajuste do modelo estatístico linear generalizado; o Nível de significância (ρ), que representa a probabilidade de rejeição da hipótese nula quando ela é verdadeira; o Desvio padrão(σ)que é uma medida que expressa o grau de dispersão de um conjunto de dados; Variância (σ^2) que mostra a medida da distância de cada valor do conjunto até o valor médio; e o Interquartil (box plot)que faz uma análise do grau de dispersão ao redor da medida da centralidade dos dados.

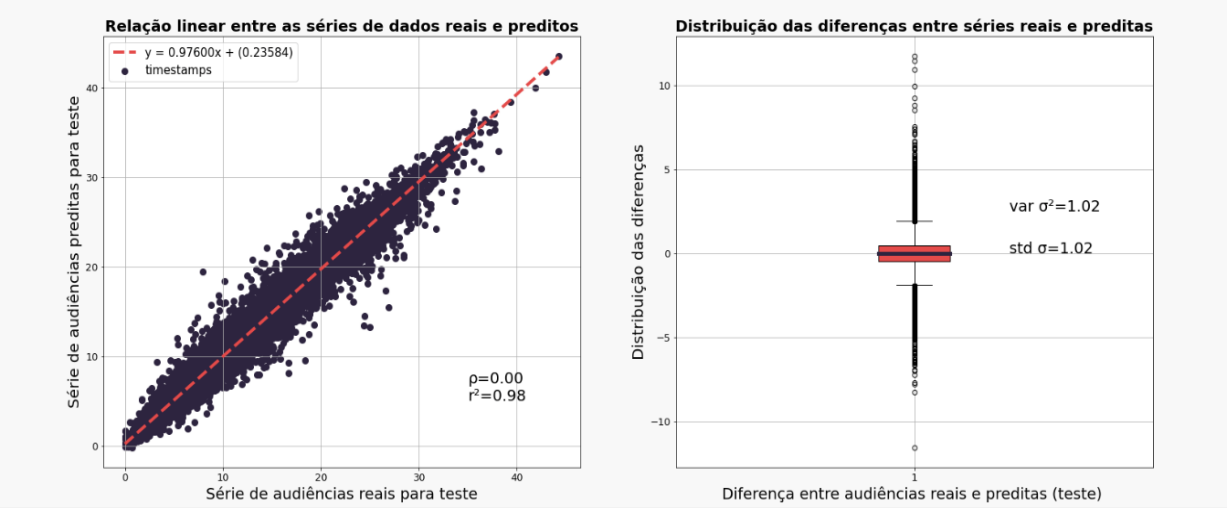
Além disso, outro ponto importante a ser ressaltado é a diferença entre os Dados de Treino e os Dados de Teste. Os dados de treino são aqueles apresentados ao algoritmo de Machine Learning para a criação do modelo. Por outro lado, os dados de teste são aqueles apresentados ao modelo após sua criação, simulando previsões reais e permitindo que o resultado final seja analisado.

4.5.1. Resultados LGBM

Resultados Treino



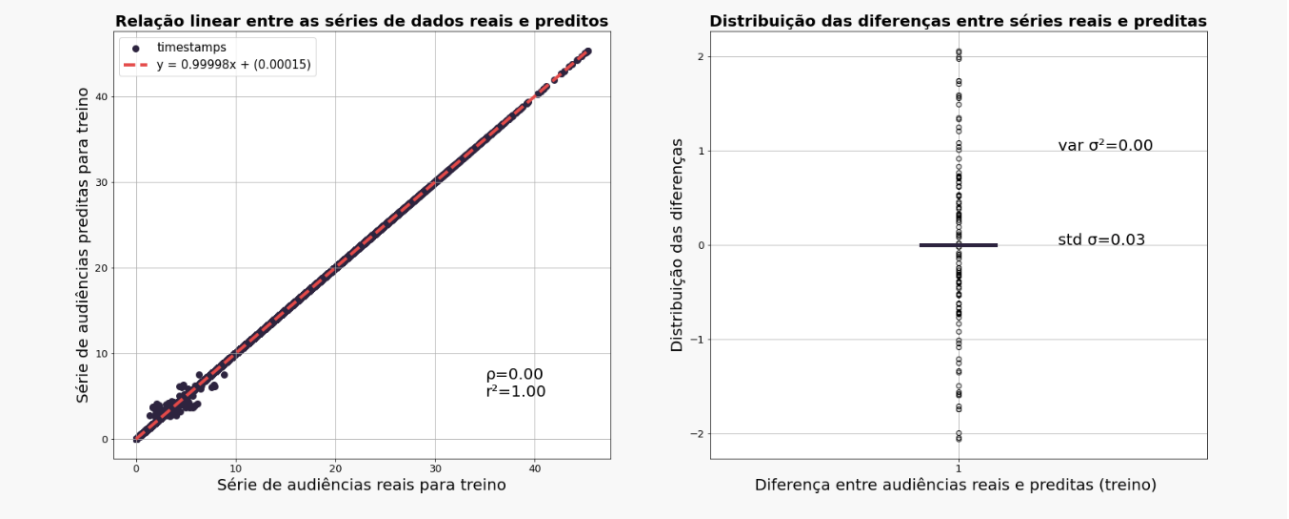
Resultados Teste



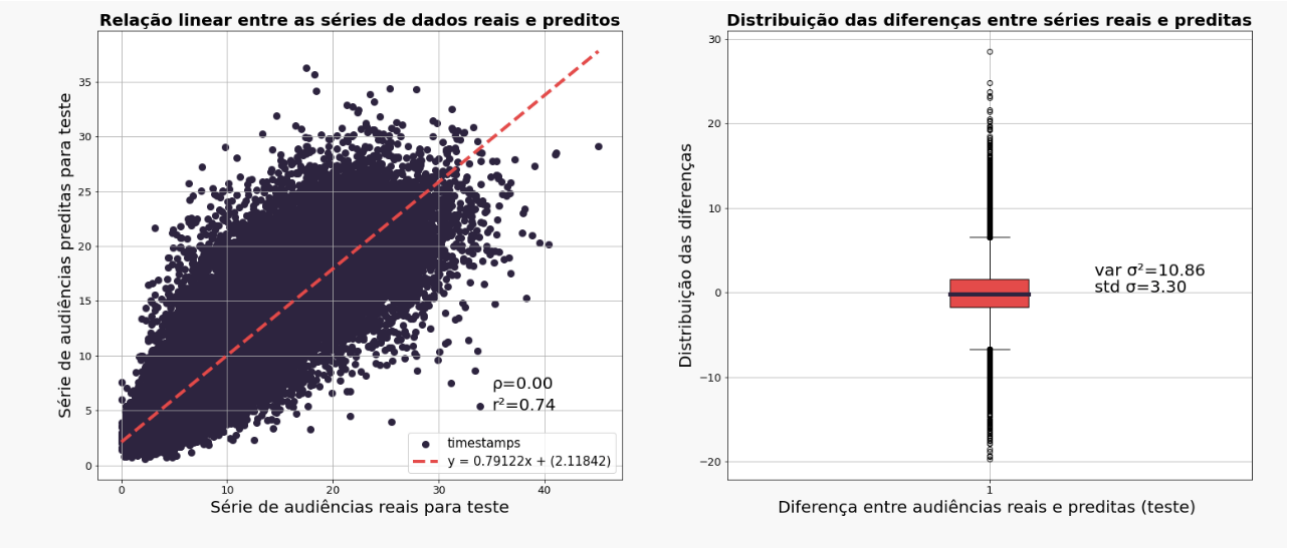
Os resultados dos gráficos acima mostram que o treino do modelo LGBM apresenta um coeficiente de determinação igual a 1, uma variância e um desvio padrão de 0,18. Além disso, não possui nível de significância. Em contrapartida, os resultados de teste mostram que o coeficiente de significância diminuiu, e passou a valer 0,98, enquanto a variância e o desvio padrão aumentaram para 1,02.

4.5.2. Resultados KNN

Resultados Treino



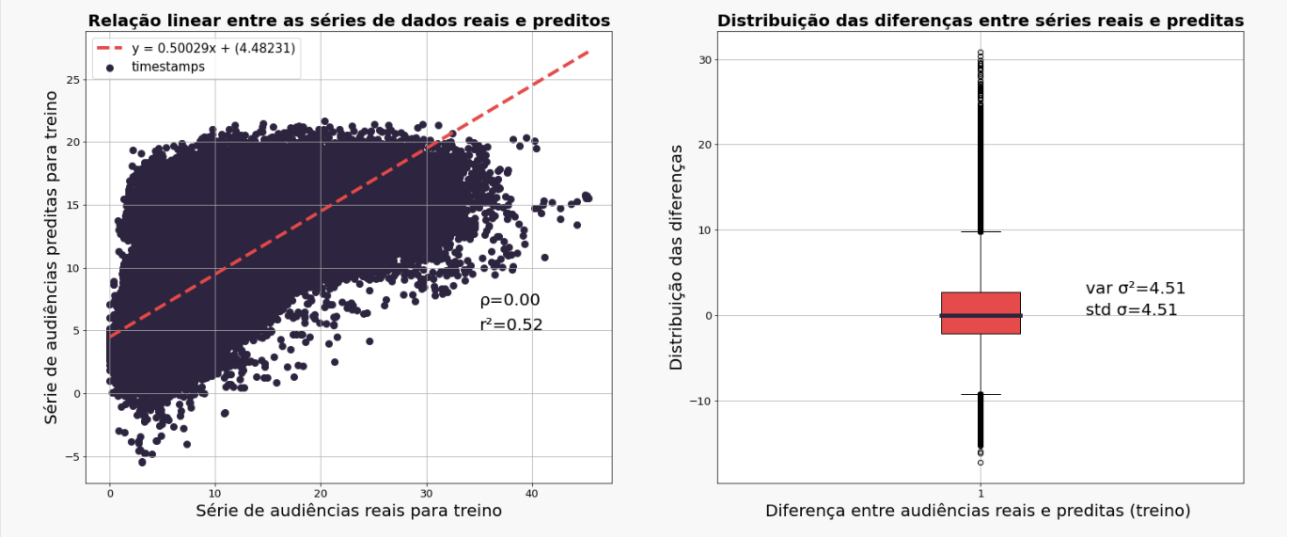
Resultados Teste



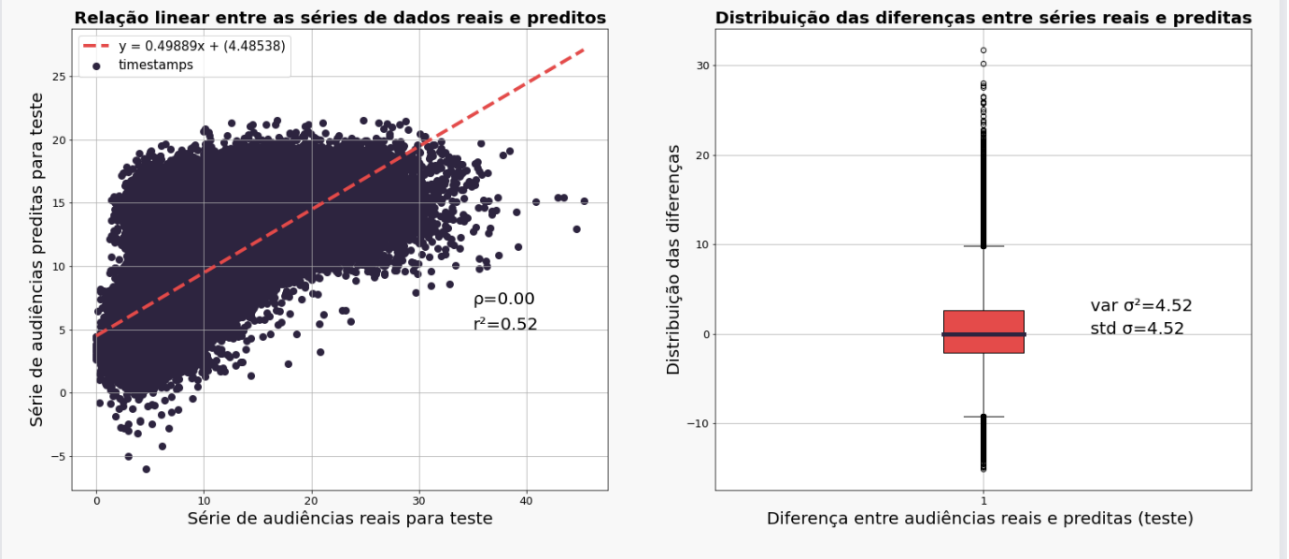
Treinando o modelo KNN(k-nearest neighbors algorithm) se obteve que o resultado do r^2(coeficiente de determinação) dos dados de treino é igual a 1, além de apresentarem uma variância e desvio padrão de 0,00 e 0,03, respectivamente. De modo geral, não possuem níveis de significância. Analisando os gráficos de teste os resultados mostram que o coeficiente de significância diminuiu 26 décimos , por conseguinte a variância e o desvio padrão aumentaram mais que o triplo, ficando com 10,86 e 3,30, respectivamente.

4.5.3. Resultados Support Vector Machine

Resultados Treino



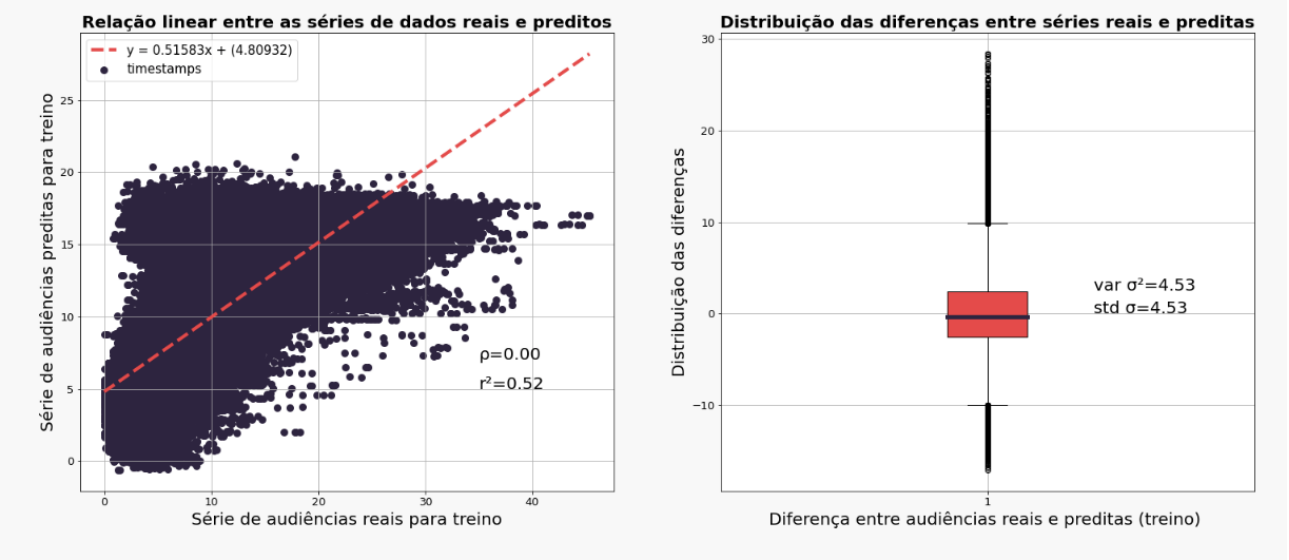
Resultados Teste



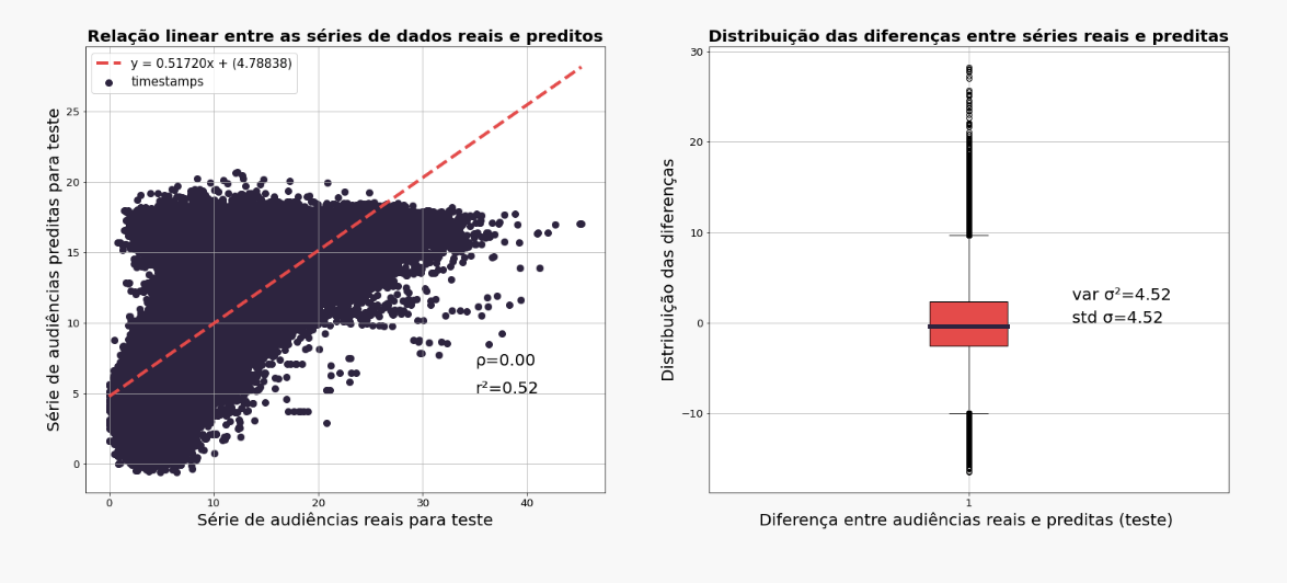
Os resultados dos gráficos de treino e teste do Support Vector mostram que tanto seu treino, quanto seu teste, obtiveram resultados muito semelhantes. Isso se deve ao fato do Support Vector ser um tipo de regressão linear, as quais normalmente conseguem resultados similares em seus treinos e testes. Além disso, tivemos um desvio padrão de 4,51 e 4,52, no treino e no teste, respectivamente, e com isso podemos ver que o modelo não teve muito sucesso.

4.5.4. Resultados Linear Regression

Resultados Treino



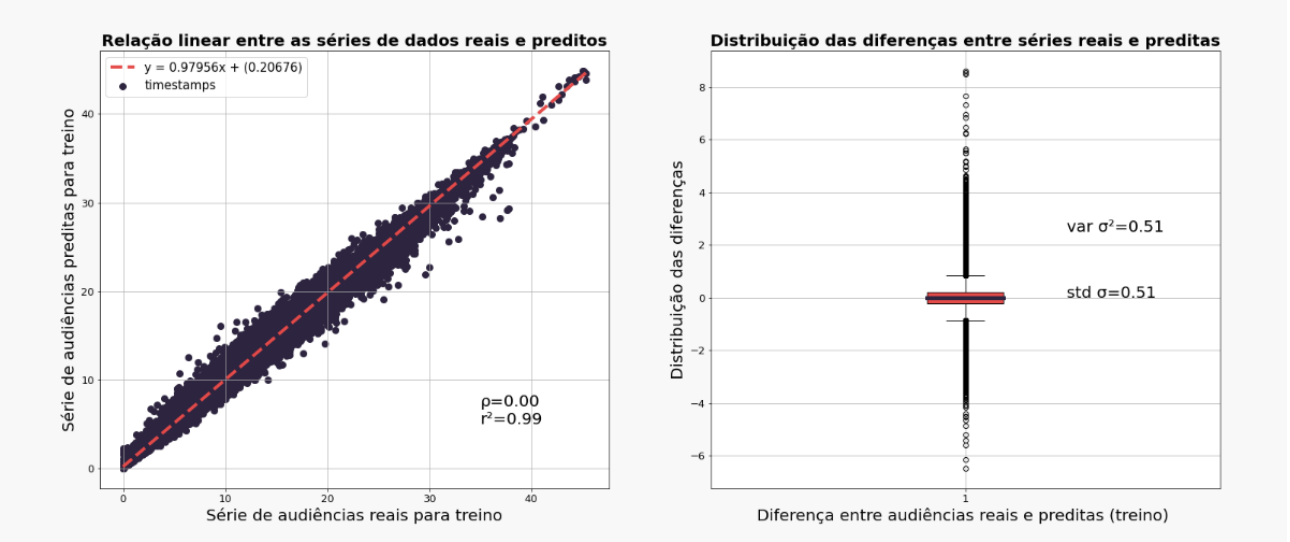
Resultados Teste



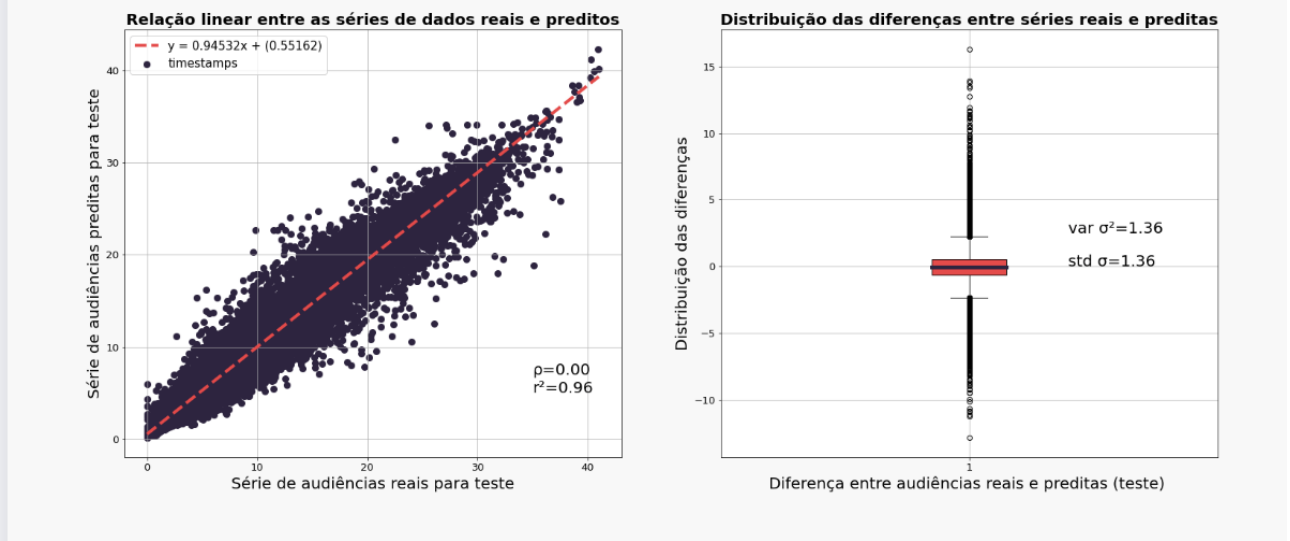
Os resultados de treinamento do modelo de regressão linear, demonstram um r^2 (coeficiente de determinação) equivalente a 0,52, também, apresentam uma variância e desvio padrão de 4,53. Já na análise dos gráficos de teste os resultados mostram que o coeficiente de significância se manteve ,enquanto a variância e o desvio padrão diminuem, passando de 4,53 para 4,52. De modo geral, tanto o treino quanto o teste tiveram resultados semelhantes, sem muito sucesso.

4.5.4. Resultados Random Forest

Resultados Treino

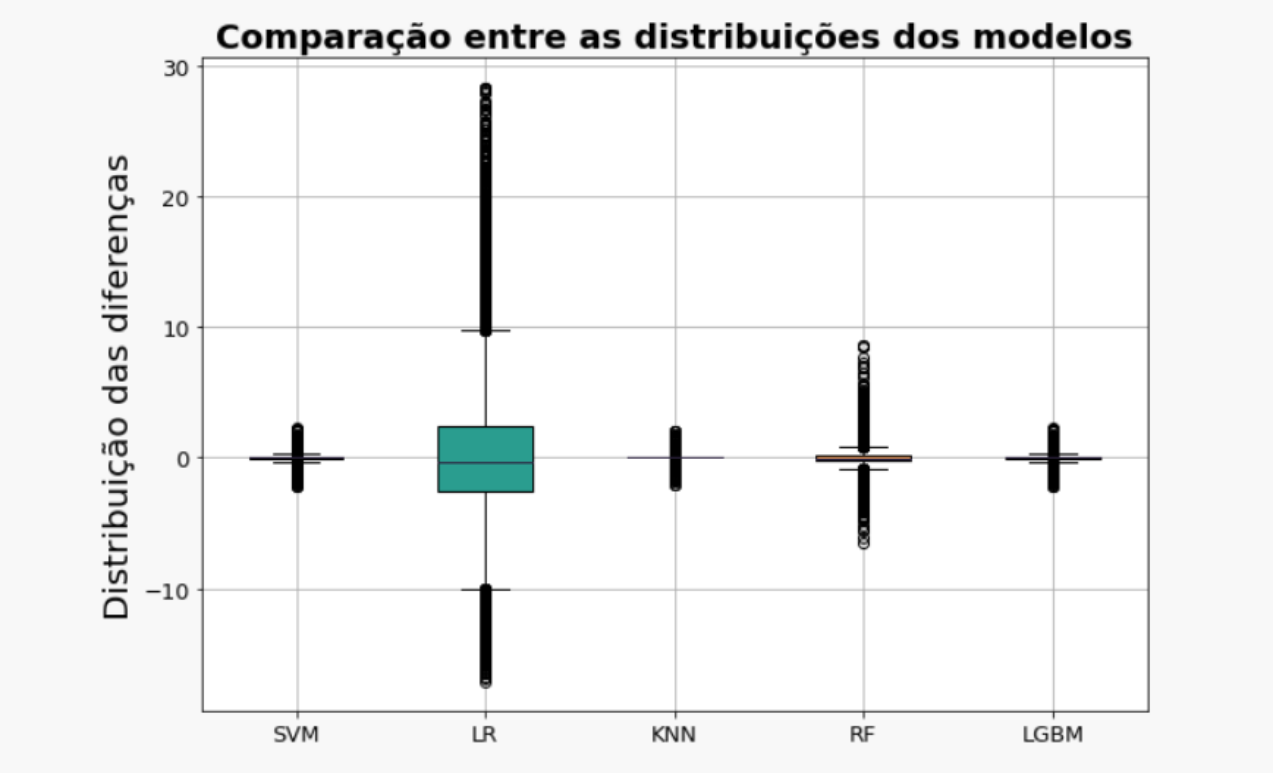


Resultados Teste



Os gráficos acima foram gerados com a intenção de analisar o modelo preditivo de Random Forest no treino e no teste. Os resultados do treino foram de um nível de significância (ρ) igual a 0,00, o r^2 (coeficiente de determinação) igual a 0,99, desvio padrão(σ) e variância(σ^2) de 0,51. Todavia, para os resultados de teste, os resultados são um nível de significância (ρ) que se mantém em 0,00, o r^2 igual a 0,96,sofrendo assim, um decréscimo de 0,03 em relação aos dados de treino, desvio padrão(σ) e variância(σ^2) se mantendo com o mesmo valor de 1,36.

4.5.5. Comparação



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.

(1) https://agenciabrasil.ebc.com.br/geral/noticia/2022-01/tv-brasil-avanca-e-ja-e-5a-emissora-mais-assistida-do-pais

(2) https://www.poder360.com.br/midia/globo-registra-receita-de-r-144-bi-e-prejuizo-de-r-173-mi/#:~:text=A%20Globo%20Comunica%C3%A7%C3%A3o%20e%20Participa%C3%A7%C3%B5es,a%20n%C3%ADveis%20anteriores%20%C3%A0%20pandemia.

(3) https://noticiasdatv.uol.com.br/noticia/mercado/em-crise-no-ibope-sbt-lucra-r-141-milhoes-gracas-futebol-e-show-do-milhao-81153

(4) https://noticiasdatv.uol.com.br/noticia/mercado/record-investe-r-622-milhoes-mas-tem-lucro-menor-que-o-sbt-em-2021-81476

(5) https://oglobo.globo.com/economia/negocios/noticia/2022/05/tv-aberta-e-canais-por-assinatura-concentram-79percent-do-consumo-de-video-do-brasileiro.ghtml

(6) https://www.uol.com.br/splash/noticias/ooops/2022/02/04/veja-o-ranking-de-ibope-da-tv-aberta-redetv-ja-ronda-o-traco.htm

Anexos

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