João Elso dos Reis Vieira

Comparação Prática de Modelos Lineares Preditivos para Métricas de Arrecadação de Concessionária de Rodovias

Goiânia

2022

SIBI
SISTEMA DE BIBLIOTECAS UFG

EMC

UFG

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC nº 1204/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG):

Nome completo do autor: João Elso dos Reis Vieira

Título do trabalho: **Comparação Prática de Modelos Lineares Preditivos para Métricas de Arrecadação de Concessionária de Rodovias**

2. Informações de acesso ao documento:

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

Independente da concordância com a disponibilização eletrônica, é imprescindível o envio do(s) arquivo(s) em formato digital PDF do TCCG.

João Elso dos Reis Vieira
Assinatura do autor[2]

Ciente e de acordo:

Weber Martins
Assinatura do orientador[2] Data: 07/04/2022

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: (a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;

[2] As assinaturas devem ser originais sendo assinadas no próprio documento, imagens coladas não serão aceitas.

João Elso dos Reis Vieira

Comparação Prática de Modelos Lineares Preditivos para Métricas de Arrecadação de Concessionária de Rodovias

Trabalho de conclusão de curso apresentado na Escola de Engenharia Elétrica, Mecânica e de Computação como requisito para a conclusão do curso de Engenharia de Computação e obtenção do título de Engenheiro de Computação.

Universidade Federal de Goiás – UFG

Escola de Engenharia Elétrica, Mecânica e de Computação (EMC)

Orientador: Prof. Dr. Weber Martins

Goiânia

2022

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Vieira, Joao Elso dos Reis
Comparação Prática de Modelos Lineares Preditivos para Métricas de Arrecadação de Concessionária de Rodovias [manuscrito] : J.E.R. Vieira, Escola de Engenharia Elétrica, Mecânica e de Computação, UFG / Joao Elso dos Reis Vieira. - 2022.
XXI, 21 f.

Orientador: Prof. Dr. Weber Martins.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Escola de Engenharia Elétrica, Mecânica e de Computação (EMC), Engenharia da Computação, Goiânia, 2022.

Inclui fotografias, tabelas.

1. armazém de dados. 2. concessionária. 3. inteligência de negócios. 4. modelos preditivos. 5. mineração de dados. I. Martins, Weber, orient. II. Título.

CDU 004

João Elso dos Reis Vieira

Comparação Prática de Modelos Lineares Preditivos para Métricas de Arrecadação de Concessionária de Rodovias

Trabalho de conclusão de curso apresentado na Escola de Engenharia Elétrica, Mecânica e de Computação como requisito para a conclusão do curso de Engenharia de Computação e obtenção do título de Engenheiro de Computação.

Trabalho aprovado. Goiânia, 07 de abril de 2022:

Prof. Dr. Weber Martins
Orientador

Prof. Dr. Thyago Carvalho Marques
Convidado 1

Prof. Dr. Rodrigo Pinto Lemos
Convidado 2

Goiânia
2022

Comparação Prática de Modelos Lineares Preditivos para Métricas de Arrecadação de Concessionária de Rodovias

J.E.R. Vieira, Escola de Engenharia Elétrica, Mecânica e de Computação, UFG

***Resumo* — A tecnologia evolui rapidamente em todos os aspectos da vida cotidiana, inclusive nas empresas. Por causa disso, maiores quantidades de dados são geradas todos os dias, mas o conhecimento extraído deles não acompanha essa expansão. O presente trabalho introduz e compara vários modelos preditivos para geração de conhecimento para uma concessionária de rodovias, por meio de técnicas de mineração de dados e desenvolvimento de modelos preditivos, em várias etapas: extração, integração, pré-processamento, modelo preditivo propriamente dito e visualização. As métricas de arrecadação consideradas foram tráfego, equivalência de eixos e receita, agrupadas e representadas por dimensões relevantes ao negócio, sendo elas: praças de pedágio, forma de pagamento, categorias, tipo de transação, tipo de veículo e datas. Tais métricas e dimensões permitem a extração de conhecimento apresentada em relatórios, utilizando ferramenta de inteligência de negócios (Power BI). Tal ferramenta apresenta graficamente os dados, aumentando velocidade e qualidade do acesso ao conhecimento, possibilitando a análise do desempenho da empresa.**

***Termos-chave* — armazém de dados, concessionária, conhecimento, dados, dimensões, inteligência de negócios, métricas, mineração de dados, modelos preditivos.**

***Abstract*— Technology has rapidly evolved into all aspects of everyday life, including businesses. Because of this, large amounts of data are generated every day, but the amount of knowledge extracted from them does not keep the same pace. The following article seeks to demonstrate a model for generating knowledge for a highway concessionaire, using data mining techniques and the development of a predictive model. The following metrics are used: traffic, axis equivalence and revenue. They are grouped and represented by dimensions relevant to the business, namely: toll plazas, payment method, categories, transaction type, vehicle type and dates. Thus, allowing the extraction of knowledge that is presented in a report by using a business intelligence tool (Power BI). This tool presents data in graphics and visuals, increasing the speed and quality of data access, enabling the analysis of the company's performance.**

***Keywords* — Business intelligence, concessionaire, data, data mining, data warehouse, dimensions, knowledge, metrics, predictive model.**

I. Introdução

Para toda empresa é sempre importante a avaliação de seu desempenho. Portanto, é essencial avaliar o conjunto de medidas quantificáveis que registram sua performance.

2022/04/05. J.E.R.Vieira é aluno de graduação da Escola de Engenharia Eletrica Mecanica e Computação, Universidade Federal de Goiás.

Atualmente, com a evolução de computadores e sistemas integrados de gestão empresarial, é possível ter acesso fácil e rápido a grandes volumes de dados, porém não é tão trivial a sua interpretação. Por tal razão, é necessária a aplicação de técnicas para transformar esses dados em algo mais inteligível.

John Naisbitt [1] afirma, muito acertadamente, que estamos "Afogados em dados, mas famintos de conhecimento." Realmente, as empresas possuem diversos sistemas, imensos bancos de dados e diversas fontes de informação, mas apresentam grande dificuldade em integrar, tratar, visualizar e analisar esses dados, ou seja, não conseguem gerar o conhecimento suficiente para melhorar suas operações. O tratamento dos dados possibilita o acesso ao conhecimento que, consequentemente, melhora a gestão da empresa, proporcionando o desenvolvimento de estratégias e táticas com maior assertividade, gerando resultados previsíveis, replicáveis e aumentando o valor da empresa de forma consistente.

A análise de dados é ferramenta indispensável em mercados cada vez mais competitivos e dinâmicos, fazendo com que a velocidade e a conveniência de painéis de indicadores não sejam mais diferencial estratégico, mas necessidade diária para tomadores de decisão. Um painel de indicadores é a apresentação consolidada do conhecimento gerado pelos dados extraídos de diversas fontes, que possibilita análise rápida do estado atual de algumas ou todas as métricas da empresa. Quanto maior a empresa, maior e mais complexa são suas métricas, criando a necessidade de modelar e implementar soluções mais elaboradas para representá-las.

Neste trabalho, serão apresentadas técnicas e ferramentas empregadas para extrair, transformar, carregar e apresentar as principais métricas de arrecadação de uma concessionária de rodovias, responsável por 630,2 quilômetros de rodovias federais nos estados de Goiás e Minas Gerais, desde 2014.

II. Referencial Teórico

A. *Mineração de Dados*

A Mineração de Dados é uma evolução natural da tecnologia da informação, mostrada na Figura 1. A seguir, apresenta-se um breve resumo da trajetória dos bancos de dados e as ferramentas desenvolvidas:

- coleção de dados e criação de Banco de Dados na década de 1960;
- Sistemas de Gerenciamento de Banco de Dados (SGBD) na década de 1970 e começo de 1980;
- sistemas avançados de Bancos de Dados na metade

da década de 1980;

- análise de dados avançada: Armazém de Dados e Mineração de Dados no final da década de 1980 até a atualidade;
- Banco de Dados para Web na década de 1990 até a atualidade;
- nova geração de Sistemas Integrados de Dados e Informação na atualidade e, certamente, no futuro.

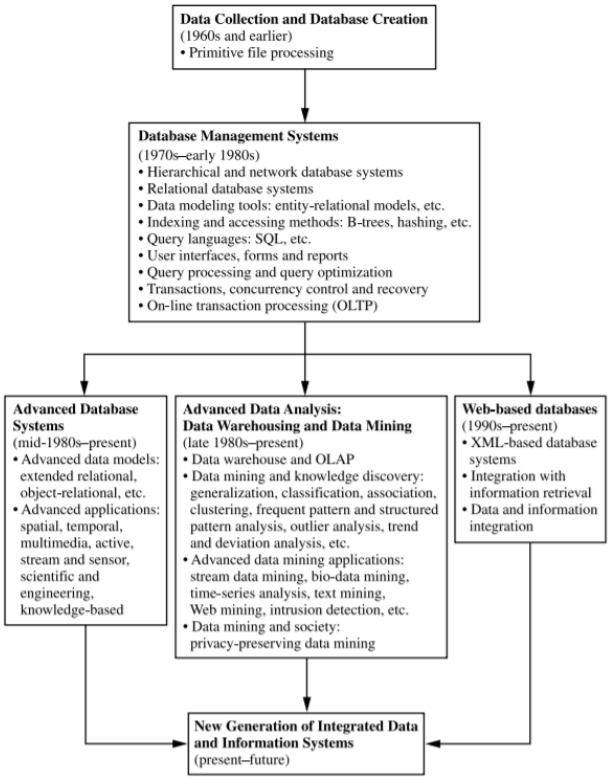

Fig. 1. Diagrama da trajetória das tecnologias de sistemas de banco de dados [2].

"Mineração de dados se refere a extração ou 'mineração' de conhecimento através de grandes volumes de dados" [2]. Para tanto, há várias ferramentas e técnicas capazes de realizar o processo de transformação de dados em conhecimentos. O processo de mineração de dados pode ser exemplificado (ver Figura 2) pelas seguintes etapas [3]:

1. Limpeza dos dados: retirada dos dados inconsistentes ou não relevantes;
2. Integração dos dados: combinação de diversas fontes de dados;
3. Seleção dos dados: busca de dados relevantes para a tarefa de análise.
4. Transformação dos dados: conversão dos dados para forma mais apropriada para mineração;
5. Mineração dos dados: aplicação de métodos inteligentes na busca de padrões;
6. Avaliação de padrões: identificação de padrões relevantes;
7. Apresentação dos dados: visualização rápida, inteligível e consistente de conhecimentos minerados.

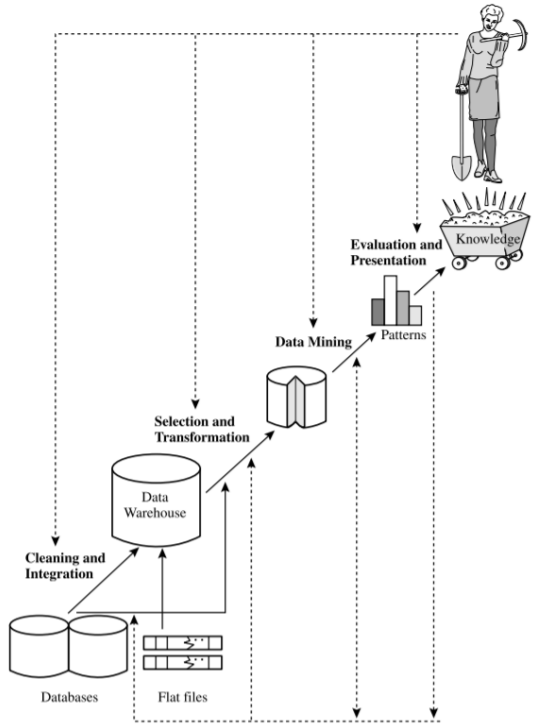

Fig. 2. Diagrama da Mineração de Dados [3].

Outra abordagem de mineração (ver Figura 3) adota as seguintes etapas [4]:

1. Coleta dos dados: integração das diversas fontes de dados;
2. Pré-processamento:
    a. seleção de características (atributos);
    b. redução de dimensão;
    c. normalização;
    d. clusterização de dados.
3. Mineração dos Dados;
4. Pós-processamento:
    a. filtragem de padrões;
    b. visualização;
    c. identificação de padrões.
5. Conhecimento.

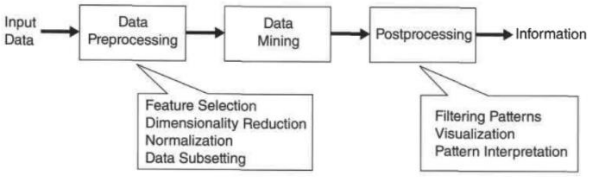

Fig. 3. Diagrama da mineração de dados [4].

Em relação as métricas de arrecadação da concessionária, utilizou-se o seguinte processo de mineração (ver Figura 4):

1. Definição do problema: descrever e definir o escopo do problema;
2. Coleta de dados: realizar a integração das diversas fontes de dados;
3. Limpeza dos dados: realizar as transformações e exclusões de dados, retirando outliers, inconsistências e informações irrelevantes;

4. Modelagem do método: elaborar o método que se adequa ao objetivo definido;
5. Validação do método: validar se o método atende ao objetivo especificado;
6. Implementação do método: sustentado em alguma(s) linguagem(ns), desenvolver o método e apresentar o conhecimento adquirido.

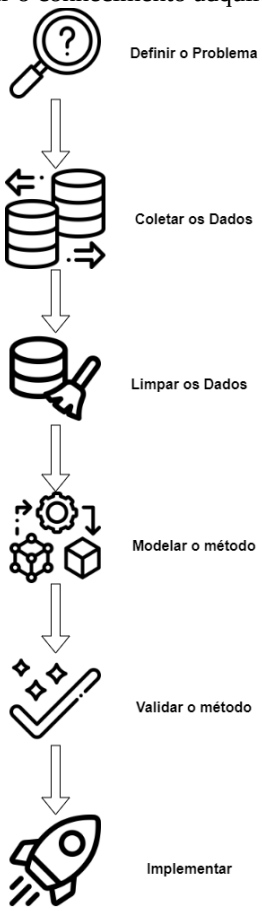

Fig. 4. Diagrama de desenvolvimento de um modelo preditivo.

Como se pode observar, o processo utilizado é um caso particular da metodologia geral, focado em buscar um objetivo concreto de forma mais prática, utilizando os métodos de Mineração de Dados.

B. *Regressão Linear*

Regressão é uma técnica para quantificar e inferir os relacionamentos entre as variáveis presente no conjunto de dados que podem ser divididas entre:

- variável dependente: Medida que dependerá do valor de outra variável, também chamada de resultado ou resposta.
- variáveis independentes: Medida que não depende de nenhuma outra variável. Definidas também como preditoras ou variáveis explanatórias.

A regressão linear é a forma mais simples e comum para análise preditiva, nela se examina o relacionamento das variáveis de forma linear (1):

$$Y_i = \beta_0 + \beta_1 X_i \quad (1)$$

Onde:

Y_i é a variável dependente;

β_0 é a constante ou intercepto;

β_1 é o coeficiente da regressão;

X_i é a variável independente.

Para o caso de mais de uma variável independente podemos reescrever a fórmula (2):

$$y(\beta, x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \quad (2)$$

O erro quadrático médio (*mean squared error* – *MSE*) é a função de custo (3) mais comum para modelos de regressão linear. Ela calcula a diferença entre os resultados previstos e o resultado real ao quadrado e posteriormente calcula a média desses valores.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_i x_i))^2 \quad (3)$$

O objetivo do modelo é ir minimizando o custo através do ajuste dos coeficientes (4):

$$\min_{\beta} \|X\beta - y\|_2^2 \quad (4)$$

C. *Viés e variância*

Viés (*bias*) pode ser definido pela diferença entre o valor correto e o valor previsto pelo modelo, ou seja, a capacidade do modelo de inferir as relações entre as variáveis. Modelos altamente enviesados possuem uma direção pré-concebida forte demais sobre a sua definição de resposta, ignoram os valores do conjunto de dados ocasionando subajuste (*underfitting* ver Figura 5).

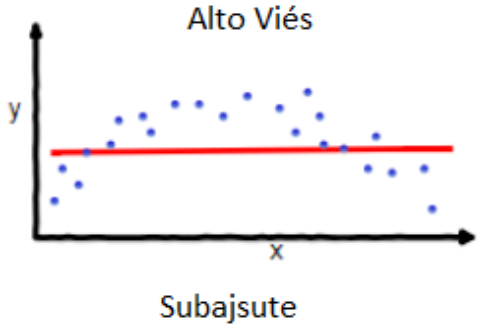

Fig. 5. Gráfico representando um modelo com subajuste.

Espera-se que os resultados sejam semelhantes para um modelo independente do conjunto de dados utilizado. Quando isso não ocorre tem-se a indicação que o modelo absorveu exageradamente os detalhes contidos no conjunto de treinamento. A variância, no contexto de aprendizado de máquina, é um tipo de erro que ocorre devido a sensibilidade do modelo a pequenas flutuações no conjunto de treinamento. Se o modelo tem alta variância, ele não é capaz de generalizar a previsão, inferindo os erros contidos nos dados, chamado de superajuste (*overfitting* ver Figura 6) [5].

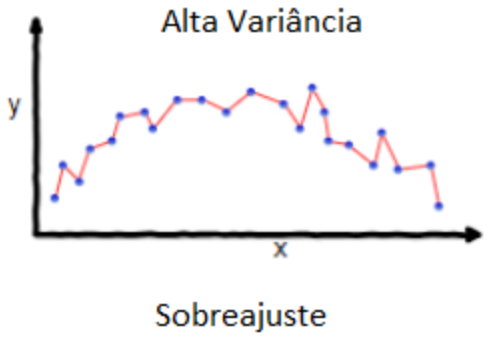

Fig. 6. Gráfico representando um modelo com sobreajuste.

Na Figura 7 é possível ver de forma gráfica os concentos de viés e variância, no qual o centro do alvo define um modelo que prevê de forma correta os valores. Quanto mais longe do centro as predições vão se tornando piores.

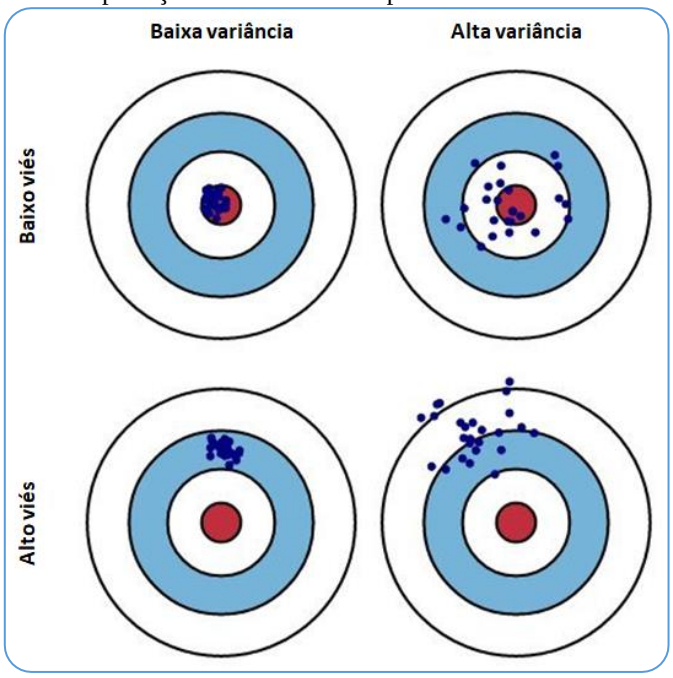

Fig. 7. Gráfico demonstrando variância e viés. [5]

Consideremos a variável que desejamos prever como Y, e as variáveis independentes, X, tal que (5):

$$\mathrm{Y} = f(X) + \varepsilon \quad (5)$$

onde o termo de erro ε é normalmente distribuído com a média zero $\varepsilon \sim N(0, \sigma_e)$.

Com erro quadrático médio é possível derivar a expressão para o erro esperado de uma regressão $\hat{f}(x)$ com entrada $X = x_0$ [6]:

$$Err(x) = E\left[\left(Y - \hat{f}(x)\right)^2\right] \quad (6)$$

$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + \sigma_e^2 \quad (7)$$

$$Err(x) = Viés^2 + Variância + Erro\ \text{irredutível} \quad (8)$$

Trabalhar com viés e variância em sua essência é lidar com subajuste e superajuste. Quantos mais parâmetros são adicionados ao modelo, maior sua complexidade e consequentemente a variância se torna o principal problema enquanto o viés cai.

O ponto ideal de complexidade do modelo é quando a diminuição no viés é equivalente a uma redução na variância (ver Figura 8).

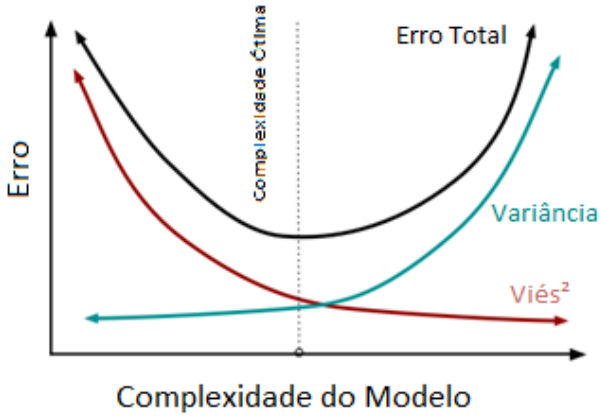

Fig. 8. Gráfico demonstrando variância e viés. [5]

Matematicamente expressado como:

$$\frac{dVies}{dComplexidade} = -\frac{dVariancia}{dComplexidade} \quad (9)$$

Quando se passa esse ponto estamos efetivamente superajustando o modelo, e para os casos em que a complexidade é inferior a ele o modelo sofre subajuste.

D. Regressão Polinomial

Regressão polinomial pode ser definida como uma forma de regressão na qual a potência máxima da variável independente é maior que 1. Sendo assim, a linha de melhor resultado não se aproxima a uma reta, mas a uma curva. Isso se aplica bem quando a relação entre a variável resposta e a variável preditora não é necessariamente linear.

Sua fórmula matemática geral (10):

$$Y = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \cdots + \beta_p x_j^p + \varepsilon \quad (10)$$

onde:
Y é a variável dependente;
β_0 é a constante ou intercepto;
β_i são coeficientes do modelo;
p é o grau do polinômio;
ε é o erro do modelo.

Dependendo do grau utilizado pelo modelo, pode-se sofrer problemas de superajuste ou subajuste conforme Figuras 9 e 10. Quando definido o grau correto o modelo se aproxima melhor e de forma mais genérica da linha de melhor resultado conforme Figura 11 [7].

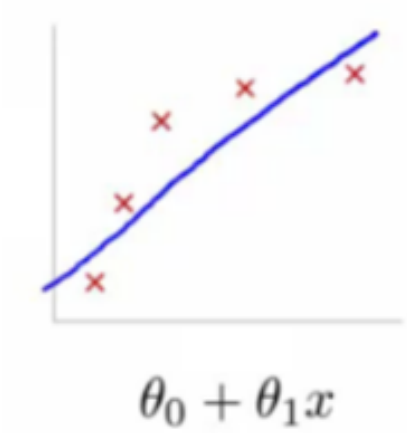

Fig. 9. Gráfico regressão polinomial de grau 1, com subajuste [7].

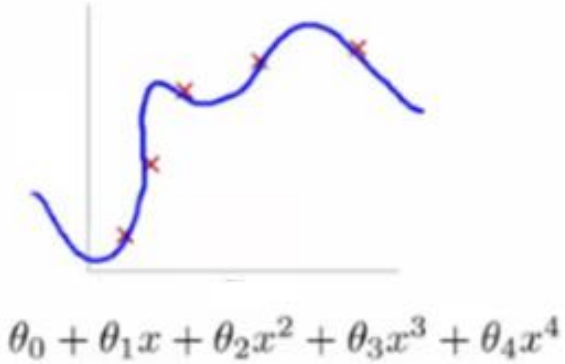

Fig. 10. Gráfico regressão polinomial de grau 4, com sobreajuste [7].

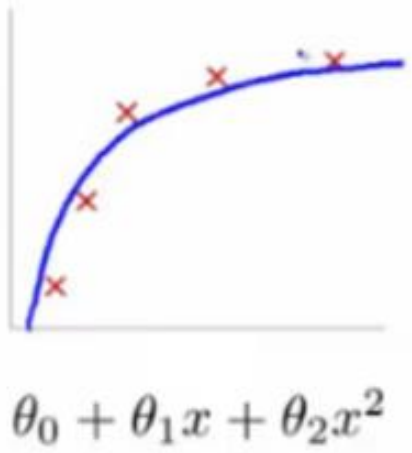

Fig. 11. Gráfico regressão polinomial de grau 2, com ajuste mais assertivo [7].

E. Regressão de Ridge e Lasso

As condições para um problema ser bem-condicionado, conhecidas como condições de Hadamard são [8]:

- existência: existe uma solução;
- unicidade: a solução é única;
- continuidade ou estabilidade: a solução possui uma dependência contínua com os dados de entrada, ou seja, pequenas mudanças nas condições iniciais e de contorno causam alterações pequenas na solução.

Se uma dessas condições não for satisfeita, o problema é considerado mal condicionado. Os problemas do mundo real geralmente são mal condicionados (*ill—posed*).

Problemas de regressão linear tentam aproximar sua função, utilizando um conjunto de dados. Por causa disso, são problemas mal condicionados. As regiões nas quais não existem dados disponíveis tornam o mapeamento de entrada-saída da função incompleto, ou seja, a solução não é única.

Com a intenção de tornar o problema bem condicionado, pode-se incluir conhecimento prévio sobre ele. Se não houver o conhecimento a solução deve ser suave o suficiente para reconstruir de forma única o mapeamento do domínio e contradomínio da função através de um conjunto de amostras insuficiente [9].

Na regressão de Ridge é inserida informação *a priori* por meio de uma penalidade. Ela cumpre o papel de diminuição dos coeficientes, que consequentemente reduzem a complexidade do modelo. Ela também é conhecida como Regularização L2. Essa penalidade serve para tratar o problema de multicolinearidade, no qual as variáveis independentes possuem relações lineares exatas ou aproximadamente exatas.

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \quad (11)$$

Para a regressão de LASSO (*Least Absolute Shrinkage and Selection Operator*) a penalidade inserida trabalho com valores absolutos dos coeficientes, possibilitando assim a remoção de parâmetros do modelo. Ela é muito utilizada para problemas com muitas variáveis independentes. Essa informação a priori também é conhecida como Regularização L1.

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \quad (12)$$

No geral a regularização tenta tratar o problema de sobreajuste do modelo, adicionando uma penalidade a função custo, de tal forma a suavizar a função e diminuir a variância da solução.

III. PROCESSO

A. Definição do Problema

O problema principal a ser resolvido envolve as principais métricas de arrecadação da concessionária de rodovias, sendo elas:

1. Tráfego: quantidade de veículos que passaram pela concessão;
2. Eixo equivalente: um veículo pode ter vários eixos e sua classificação é feita pela sua quantidade de eixos, resultando no valor da tarifa a ser paga. Por isso,

deve-se também saber a quantidade de eixos equivalentes trafegados;
3. Receita: valor de fato alcançado pela cobrança de pedágio.

As métricas devem, necessariamente, ser representadas em conjuntos que sejam relevantes para a geração de conhecimento e, assim, auxiliar na gestão e tomada de decisão. Esses conjuntos foram definidos pelas dimensões:

1. Praças de pedágio: os locais nos quais são feitas as transações e passagem dos veículos.
2. Categorias: classificação por eixos dos veículos. Atualmente são 15 categorias diferentes.
3. Forma de pagamento: são as formas de pagamento utilizadas para pagar a tarifa.
4. Tipo de transação: os veículos possuem duas opções quando passam em uma praça de pedágio: pagamento automático e pagamento manual.
5. Tipo de veículo: os veículos podem ser classificados como leve ou pesado, sendo veículos leves os de passeio (carros e motos) e os pesados os veículos comerciais (caminhões).
6. Data: visualização das métricas por período, para a solução ano e mês.

Portanto é possível perceber que para a visualização e análise de dados as métricas representam os valores quantitativos e as dimensões representam os valores qualitativos. Uma vez determinado o problema, os dados foram trabalhados para ser alcançado o conhecimento, apresentando-o de forma simples e explicita.

B. Coletar os Dados

Os dados estavam distribuídos em diversas fontes distintas, devendo-se realizar a integração desses dados. Para isso foi desenvolvida a seguinte arquitetura (ver Figura 12):

- Servidor de Integração e Transformação
  - O serviço de integração é feito pelo Microsoft SQL Server Integration Services (SSIS)
  - Os serviços de transformação são feitos por *scripts Python*.
- Armazém de Dados
  - Banco de dados *Staging* no qual é feito a carga inicial dos dados e utilizado para realizar transformações e integrações.
  - Banco de dados Armazém no qual é armazenado as tabelas e dados já processados, utilizado para alimentar o software de apresentação.

Como existem várias fontes de dados se torna inevitável o uso de um serviço de integração. Desse modo, foi escolhido para a arquitetura o Microsoft SQL Integration Services, no qual é feito todo o controle de comunicação, transporte, carga, notificação e registro das transações. Através desse serviço é possível chamar os *scripts Python* para também realizar as transformações necessárias.

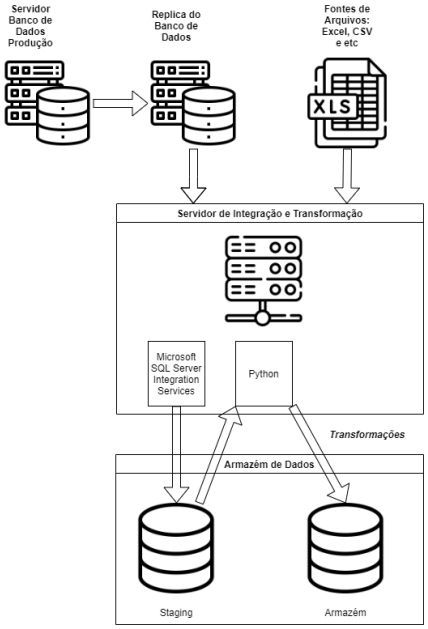

Fig. 12. Arquitetura de coleta dos dados.

Outro ponto que deve ser considerado é a performance e o impacto no servidor do banco de dados de produção. Não é recomendado que as extrações sejam feitas diretamente nele, podendo comprometer o funcionamento da aplicação de arrecadação da concessionária. Por causa disso, foi criado um banco réplica, utilizando a ferramenta de espelhamento disponível no SQL Server.

Para a extração de arquivos (Excel, CSV, texto e etc) foram utilizados os conectores de fontes do SSIS (ver Figura 13):

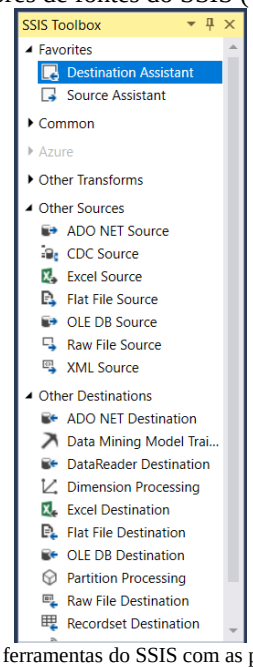

Fig. 13. Caixa de ferramentas do SSIS com as possíveis fontes.

O SSIS utiliza pacotes para a construção dos serviços de integração com as fontes de dados e o armazém de dados. Esses pacotes depois são implementados e executados via SQL Agent no SQL Server do Armazém de Dados (ver Figura 14).

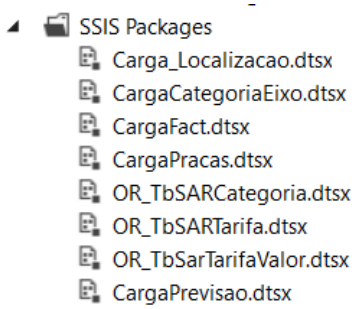

Fig. 14. Pacotes de extração do SSIS.

1. Carga Localização

Foi feita a carga de uma planilha Excel com as informações de localização geográfica das praças de pedágio (ver Figura 15).

Fig. 15. Fluxo de dados do SSIS para a extração da tabela de localizações geográficas das praças de pedágio.

2. Carga Categoria X Eixo

Foi feita a carga de uma planilha Excel com as informações de relação entre a categoria do veículo e o número de eixos equivalente (ver Figura 16).

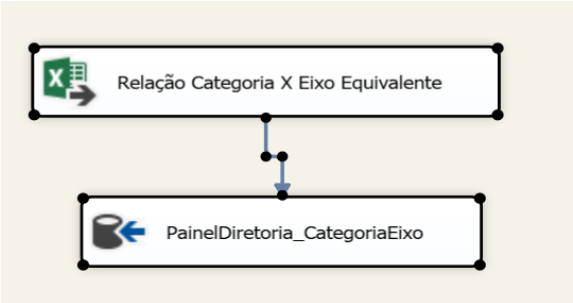

Fig. 16. Fluxo de dados do SSIS para a extração da tabela de relação entre categorias e eixos equivalente.

3. Carga Tarifas

Foi feita a carga das tabelas de Tarifas e Valor Tarifa do banco de dados réplica para o banco de dados *Staging* (ver Figura 17).

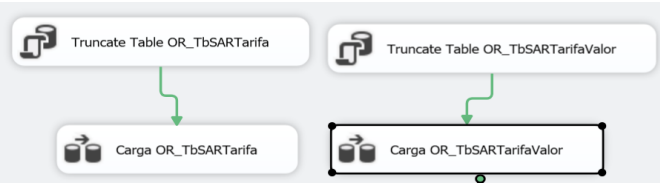

Fig. 17. Fluxo de dados do SSIS para a extração das tabelas de tarifas e valores de tarifas.

4. Carga Tabela Fato

Para a tabela de fatos, ou seja, a tabela que contém as métricas da solução, foi feita a extração através de uma consulta SQL que realiza os agrupamentos necessários para as dimensões escolhidas (ver Figura 18). A consulta SQL será apresentada com mais detalhes no tópico "*C. Limpeza dos Dados"*.

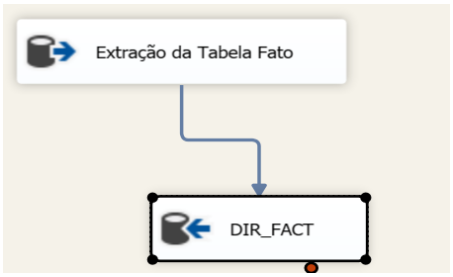

Fig. 18. Fluxo de dados do SSIS para a extração da tabela fato que contém as métricas.

5. Carga Previsão

O fluxo de dados da carga de previsão é mais complexo pois recomenda-se preparar todas as tabelas para o modelo de previsão, calcular a previsão, transformar ela no formato da tabela fato e realizar a atualização dos dados na tabela fato (ver Figura 19).

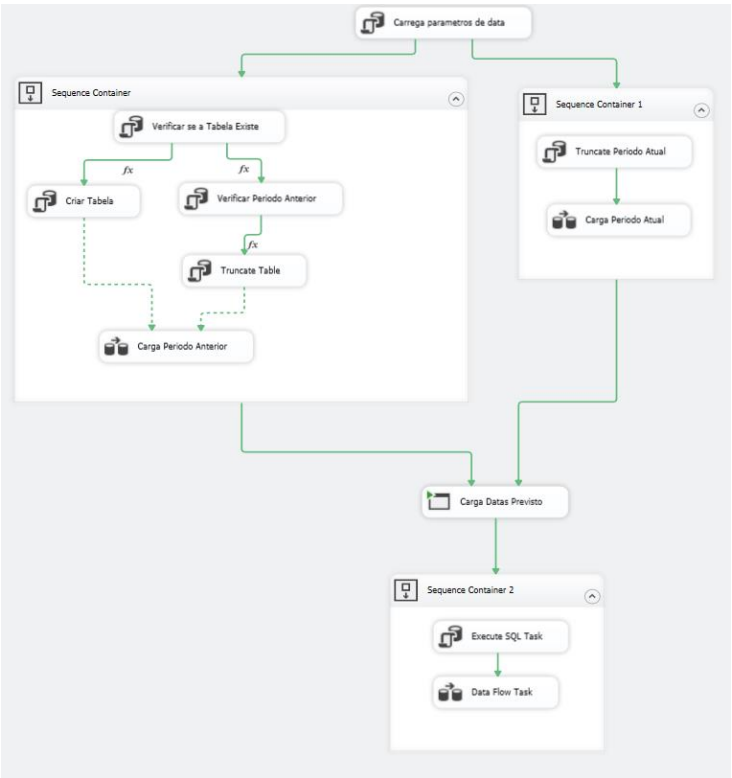

Fig. 19. Fluxo de dados do SSIS para a execução da previsão.

6. Carga Conjunto de Dados (*dataset*)

Foi extraído do banco de replicação o conjunto de dados, no formato de tabela chamada Temp_DataSet, que será utilizado para o treinamento dos modelos que utilizam aprendizado de máquina. Ele é composto pelas variáveis conforme Tabela 1.

Independentes	Dependente
Data	Tráfego
Ano	
Mês	
Dia	
Dia da semana	
Praça	
Categoria	
Feriado	

Tab. 1. Variáveis dependentes e independentes.

C. *Limpar os Dados*

Atualmente, as bases de dados de produção estão suscetíveis a dados inconsistentes ou faltantes. Isso ocorre por causa das suas diversas fontes (sistemas, arquivos, web e etc) e erros de entrada. Dados de baixa qualidade acabarão resultando em uma mineração inadequada [10]. Sendo assim, deve-se limpar os dados em três passos:

1. Limpeza dos Dados: remover as inconsistências e anomalias. Para a mineração das métricas de arrecadação foram identificadas e executadas as seguintes correções:
    - Entre 21 e 30 de maio de 2018 houve a greve dos caminhoneiros que acabou alterando significativamente as métricas da rodovia, de tal forma que foi preciso corrigir as informações substituindo-as pela média de cada dia da semana, tendo em vista o comportamento cíclico dos dias da semana na rodovia. *Exemplo: dia 21 de maio foi uma segunda-feira, seu valor foi substituído pela média dos valores das segundas-feiras contidas entre 01 e 20 de maio).*
    - Nos feriados ocorre um comportamento anormal nas rodovias aumentando muito seu tráfego. Por causa disso, foi criado uma tabela com os feriados até o ano de 2099, ela é utilizada para identificar os feriados e substituir os valores de tráfego pela média do dia da semana.
    - Pandemia de 2020, nos primeiros meses houve uma mudança brusca no tráfego, os valores referentes a esses meses (maio de 2020 até novembro de 2020) foram substituídos pela média dos meses considerando os anos anteriores.
    - Remoção das categorias de erro (categoria 0) que ocorre quando não possível identificar a categoria do veículo, ou seja, os sensores não foram capazes de classificar a categoria do veículo.
2. Integração dos Dados: unificar e juntar os diversos dados relacionados. Para isso foram montados dois conjuntos de tabelas: as tabelas de dimensões e a tabela de fato. A unificação e agregação dos valores são feitos conforme as dimensões já descritas anteriormente. Utiliza-se o SSIS para executar uma consulta SQL que cria a tabela fato do modelo, ela realiza os agrupamentos.

   As dimensões: Forma de Pagamento e Tipo de Transação, foram unificadas no campo FormaPgto da tabela fato para reduzir o número de linhas da tabela. Dentro do software de apresentação elas são separadas novamente (ver Tabela 2).

Dimensão	Campo da Tabela Fato
Praças	Praça
Categorias	Categoria
Forma de pagamento	FormaPgto
Tipo de transação	FormaPgto
Tipo de veículo	TipoVeiculo
Data	AnoMes

Tab. 2. Tabela da relação entre os campos da tabela fato com as dimensões do problema.

3. Transformação dos Dados para um formato mais fácil de ser processado e apresentado. Existem no cadastro do sistema de arrecadação da empresa 78 formas de pagamentos distintos que não são tão relevantes para as dimensões que se deseja utilizar. Dessa forma, foi criado uma *Stored Procedure* que transforma as formas de pagamento em grupos de pagamento, que são mais fáceis de serem filtrados e apresentados.

D. *Modelar o método*

O modelo foi pensando em três pontos principais: o passado, o presente e o futuro.

Para o passado a intenção é disponibilizar os dados de tal forma que possibilite a análise descritiva dos dados. O que auxilia a responder perguntas sobre o que aconteceu com base em dados históricos. Tal como possibilitar uma análise de diagnóstico, explicando por que certos eventos aconteceram, identificando anomalias nos dados e tentando descobrir as relações que expliquem suas causas.

Em relação ao presente, pretende-se permitir análise prescritiva indicando quais ações e estratégias devem ser elaboradas para alcançar os objetivos e metas da empresa.

Para o futuro foram propostos métodos preditivos para calcular e prever o que acontecerá no futuro, a fim de comprar qual apresenta o melhor resultado.

O passado e o presente foram representados através de um relatório utilizando o software da Microsoft de inteligência de negócios, Power BI. Eles são organizados em diversos gráficos e representações que possibilitam as análises necessárias.

Para o futuro foram utilizados os seguintes métodos com o objetivo de prever o mês de março de 2022:

1. Modelo empírico
2. Médias dos valores de tráfego
3. Regressão Linear
4. Regressão Polinomial
5. Regressão de Ridge
6. Regressão LASSO

1. Modelo empírico

Um método preditivo baseado no histórico dos dados que apresentam algumas características interessantes:

- o tráfego de veículos segue um padrão semanal, no qual datas com o mesmo dia da semana possuem tráfego semelhante.
- existe uma sazonalidade referente ao tráfego, um exemplo, é o aumento do tráfego em períodos de safra.

Considerando essas características foi elaborado o modelo utilizando o período do ano anterior (que inclui a sazonalidade) somado a uma tendência atual do tráfego.

Para realizar os cálculos, primeiramente foram extraídas duas tabelas (ver Figura 20) com os dados de tráfego, o mês corrente (que se deseja prever) e o mês referência (que é o mesmo mês de um ano que será utilizado como base para o cálculo). No modelo proposto utilizamos o ano anterior. As duas tabelas foram então inseridas no armazém de dados, que posteriormente serão processadas por *script Python* que gerará as tendências, calculadas por dia, praça de pedágio e categoria. Dessa forma, gera-se uma tabela com essas tendências, ela também é inserida no armazém de dados. Em seguida são calculados os dias e os valores de tráfego usando as tendências e os valores do mês anterior, de tal forma que se monta outra tabela com o mês previsto. A última etapa consiste em utilizar a tabela com os valores de tráfego calculados e as demais tabelas de dimensões para preencher a tabela fato com os dados encontrados. Como somente foi calculado o tráfego por categoria e dia, é necessário determinar o valor da receita (utilizando os valores das tarifas) e o eixo equivalente (utilizando a tabela categoria × eixo equivalente), e realizar o agrupamento das métricas.

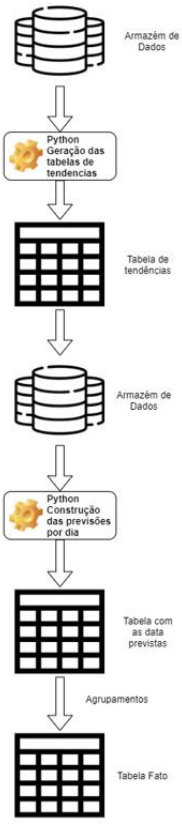

Fig. 20. Método para previsão do tráfego.

Geralmente, duas datas iguais em anos diferentes estão em dias da semana distintos, ou seja, não há equivalência entre o dia da semana e a data. Por causa disso, é necessário calcular um deslocamento que ajusta o dia da semana com a data mais próxima (no ano referência) que possuí o mesmo dia da semana (ver Figura 21).

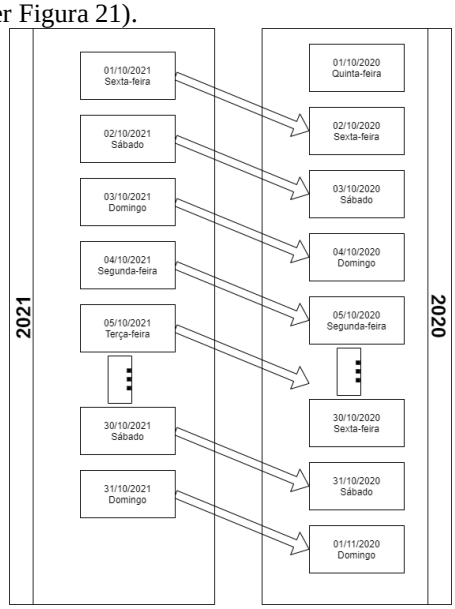

Fig. 21. Relação do dia do mês e dia da semana entre o mês atual e mês de referência.

Foi desenvolvido *Stored Procedure* que faz esse cálculo para o qual são passados como parâmetros: a data e o número de períodos (em anos).

Para calcular a tendência é utilizado a seguinte fórmula (13):

$$Tend = (T_A/T_R) - 1 \quad (13)$$

$Tend$ é a tendência,
T_A é o tráfego na data atual,
T_R é o tráfego na data equivalente no mês referência.

Esse cálculo é feito através de um *script Python* que itera os dias possíveis para o mês atual, por praça de pedágio e por categoria gerando e carregando os valores para o armazém de dados na tabela de tendências (ver Figura 22) que conterá todos os dias do mês (que já possuem tráfego), praça de pedágio e categoria.

Data	Praca	Categoria	Trafego Atual	Data Anterior	Trafego Anterior	Tendencia
2021-10-01	1	1	13241	2020-10-02	13704	-0,0337857559836544
2021-10-01	1	2	1546	2020-10-02	1576	-0,0190355329949239
2021-10-01	1	3	190	2020-10-02	206	-0,0776699029126213
2021-10-01	1	4	992	2020-10-02	1046	-0,0516252390057361
2021-10-01	1	5	31	2020-10-02	19	0,631578947368421
2021-10-01	1	6	662	2020-10-02	698	-0,0515759312320917
2021-10-01	1	7	354	2020-10-02	323	0,0959752321981424
2021-10-01	1	8	659	2020-10-02	552	0,193840579710145
2021-10-01	1	9	276	2020-10-02	307	-0,100977198697068
2021-10-01	1	11	228	2020-10-02	263	-0,133079847908745
2021-10-01	1	12	15	2020-10-02	7	1,14285714285714
2021-10-01	1	13	259	2020-10-02	237	0,092827004219409
2021-10-01	1	15	1	2020-10-02	0	0
2021-10-01	2	1	28601	2020-10-02	28409	0,00675842162694917
2021-10-01	2	2	2879	2020-10-02	2927	-0,0163990433891357
2021-10-01	2	3	430	2020-10-02	446	-0,0358744394618834
2021-10-01	2	4	1468	2020-10-02	1521	-0,034845496383958
2021-10-01	2	5	54	2020-10-02	30	0,8
2021-10-01	2	6	1070	2020-10-02	1095	-0,0228310502283106
2021-10-01	2	7	607	2020-10-02	598	0,0150501672240804
2021-10-01	2	8	922	2020-10-02	842	0,0950118764845607

Fig. 22. Tabela de tendências.

Após a geração da tabela de tendências é chamado outro *script Python* que realizará a construção da tabela de datas do

mês atual, complementando-a com o tráfego previsto. Para isso é feito a média das tendências dos dias, calculando assim a tendência mensal ($Tmensal$) (2), por praça de pedágio e categoria.

$$Tmensal = \sum Tend / \sum Dias \quad (14)$$

Com a tendência mensal os demais dias são iterados calculando o tráfego previsto (T_P) (3).

$$T_P = (T_R * Tmensal) + T_R \quad (15)$$

A partir do tráfego previsto também foram calculados os eixos equivalentes e a receita, utilizando as tabelas de dimensões. Em seguida, o resultado foi carregado para o armazém de dados em uma tabela (ver Figura 23) com as datas e os valores previstos.

Data	Praca	Categoria	Tipo Veiculo	Forma Pagamento	Trafego	Eixo Equivalente	Receita
2021-10-01	1	1	Leve	pagante;previsao;	13241	13241	38398,9
2021-10-01	1	2	Pesado	pagante;previsao;	1546	3092	8966,8
2021-10-01	1	3	Leve	pagante;previsao;	190	285	826,5
2021-10-01	1	4	Pesado	pagante;previsao;	992	2976	8630,4
2021-10-01	1	5	Leve	pagante;previsao;	31	62	179,8
2021-10-01	1	6	Pesado	pagante;previsao;	662	2648	7679,2
2021-10-01	1	7	Pesado	pagante;previsao;	354	1770	5133
2021-10-01	1	8	Pesado	pagante;previsao;	659	3954	11466,6
2021-10-01	1	9	Leve	pagante;previsao;	276	138	400,2
2021-10-01	1	11	Pesado	pagante;previsao;	228	1596	4628,4
2021-10-01	1	12	Pesado	pagante;previsao;	15	120	348
2021-10-01	1	13	Pesado	pagante;previsao;	259	2331	6759,9
2021-10-01	1	15	Pesado	pagante;previsao;	1	0	0
2021-10-01	2	1	Leve	pagante;previsao;	28601	28601	54341,9
2021-10-01	2	2	Pesado	pagante;previsao;	2879	5758	10940,2
2021-10-01	2	3	Leve	pagante;previsao;	430	645	1225,5
2021-10-01	2	4	Pesado	pagante;previsao;	1468	4404	8367,6
2021-10-01	2	5	Leve	pagante;previsao;	54	108	205,2
2021-10-01	2	6	Pesado	pagante;previsao;	1070	4280	8132
2021-10-01	2	7	Pesado	pagante;previsao;	607	3035	5766,5
2021-10-01	2	8	Pesado	pagante;previsao;	922	5532	10510,8

Fig. 23. Datas previstas.

2. Médias dos valores de tráfego

Todos os modelos a seguir utilizam a tabela *Temp_DataSet* como conjunto de dados principal separado em:

- treinamento: datas entre 2015-01-01 até 2022-02-28;
- teste: datas entre 2022-03-01 e 2022-03-31.

Para validação dos modelos foi construído um conjunto de dados com a previsão do mês de março de 2022 utilizando o valor das médias do tráfego. O cálculo foi feito considerando o tráfego aglutinado por praça e categoria, conforme a fórmula:

$$Media = \frac{1}{N}\sum_{i=1}^{n} Trafego\,(Praca, Categoria) \quad (16)$$

O valor da média foi inserido como valor previsto para cada praça e categoria em cada dia do mês de março.

O conjunto foi armazenado como tabela *Temp_DataSet_Media* (ver Figura 24).

Data	Ano	Mes	Dia	DiaSemana	Feriado	Praca	Categoria	Trafego
2022-03-01	2022	3	1	3	1	1	1	12573,1087662338
2022-03-01	2022	3	1	3	1	1	2	1202,1892001624
2022-03-01	2022	3	1	3	1	1	3	128,991873222267
2022-03-01	2022	3	1	3	1	1	4	853,046728971963
2022-03-01	2022	3	1	3	1	1	5	13,5122149837134
2022-03-01	2022	3	1	3	1	1	6	534,713821138211
2022-03-01	2022	3	1	3	1	1	7	262,322226737099
2022-03-01	2022	3	1	3	1	1	8	437,871951219512
2022-03-01	2022	3	1	3	1	1	9	349,201137748883
2022-03-01	2022	3	1	3	1	1	11	189,001221995927
2022-03-01	2022	3	1	3	1	1	12	5,68924640135478
2022-03-01	2022	3	1	3	1	1	13	136,689767631472
2022-03-01	2022	3	1	3	1	1	15	2,42644320297952
2022-03-01	2022	3	1	3	1	2	1	24787,4622564935
2022-03-01	2022	3	1	3	1	2	2	2192,18927701056
2022-03-01	2022	3	1	3	1	2	3	267,579203899269
2022-03-01	2022	3	1	3	1	2	4	1237,13002844372
2022-03-01	2022	3	1	3	1	2	5	23,8126777732629
2022-03-01	2022	3	1	3	1	2	6	842,802519301097
2022-03-01	2022	3	1	3	1	2	7	477,471353108493
2022-03-01	2022	3	1	3	1	2	8	680,135255889521
2022-03-01	2022	3	1	3	1	2	9	1281,70569105691
2022-03-01	2022	3	1	3	1	2	11	300,130081300813
2022-03-01	2022	3	1	3	1	2	12	11,4248258910283

Fig. 24. Tabela *Temp_DataSet_Media*.

3. Regressão Linear

O modelo de regressão linear múltipla utilizado para determinar a variável dependente é Y, as variáveis independentes são x_p e os coeficientes β_p através do conjunto de dados composto de todo o período de concessão exceto o mês de março. A função *MSE* foi utilizada como função custo do modelo.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \quad (17)$$

Os resultados foram armazenados na tabela *Temp_DataSet_LinearRegression* (ver Figura 25).

index	Data	Ano	Mes	Dia	DiaSemana	Feriado	Praca	Categoria	Trafego
0	2022-03-01	2022	3	1	3	1	1	1	13443,6931419885
31	2022-03-01	2022	3	1	3	1	1	2	522,480335435576
62	2022-03-01	2022	3	1	3	1	1	3	125,034255515433
93	2022-03-01	2022	3	1	3	1	1	4	554,854697924318
124	2022-03-01	2022	3	1	3	1	1	5	13,8540211612935
155	2022-03-01	2022	3	1	3	1	1	6	331,11484766307
186	2022-03-01	2022	3	1	3	1	1	7	166,10352031074
217	2022-03-01	2022	3	1	3	1	1	8	314,811784788264
248	2022-03-01	2022	3	1	3	1	1	9	517,195003839547
279	2022-03-01	2022	3	1	3	1	1	11	109,057160368375
310	2022-03-01	2022	3	1	3	1	1	12	3,92005510387241
341	2022-03-01	2022	3	1	3	1	1	13	125,678639209436
372	2022-03-01	2022	3	1	3	1	1	15	0,634450129988863
392	2022-03-01	2022	3	1	3	1	2	1	22086,939270914
423	2022-03-01	2022	3	1	3	1	2	2	958,14095607102
454	2022-03-01	2022	3	1	3	1	2	3	339,834332668019
485	2022-03-01	2022	3	1	3	1	2	4	704,980942733713
516	2022-03-01	2022	3	1	3	1	2	5	27,3961615165081
547	2022-03-01	2022	3	1	3	1	2	6	470,959381782712
578	2022-03-01	2022	3	1	3	1	2	7	253,412176772319
609	2022-03-01	2022	3	1	3	1	2	8	473,44358312109
640	2022-03-01	2022	3	1	3	1	2	9	1045,51128374381
671	2022-03-01	2022	3	1	3	1	2	11	200,383122755098
702	2022-03-01	2022	3	1	3	1	2	12	7,95560834400771

Fig. 25. Tabela *Temp_DataSet_LinearRegression*.

4. Regressão Polinomial

Para o modelo que se utilizou regressão polinomial o código fonte desenvolvido em *Python* recebeu como parâmetro o grau

dos polinômios das variáveis independentes. Os valores previstos foram armazenados na tabela *Temp_DataSet_PolynomialRegression* (ver Figura 26).

0	2022-03-01	2022	3	1	3	1	1	1	33873,2121353149
31	2022-03-01	2022	3	1	3	1	1	2	8091,46074604988
62	2022-03-01	2022	3	1	3	1	1	3	-132,314572811127
93	2022-03-01	2022	3	1	3	1	1	4	6040,38121497631
124	2022-03-01	2022	3	1	3	1	1	5	15,931648015976
155	2022-03-01	2022	3	1	3	1	1	6	3523,82192361355
186	2022-03-01	2022	3	1	3	1	1	7	1818,9595990181
217	2022-03-01	2022	3	1	3	1	1	8	1466,44564199448
248	2022-03-01	2022	3	1	3	1	1	9	1635,23675632477
279	2022-03-01	2022	3	1	3	1	1	11	167,418559789658
310	2022-03-01	2022	3	1	3	1	1	12	13,4773970842361
341	2022-03-01	2022	3	1	3	1	1	13	320,222083568573
372	2022-03-01	2022	3	1	3	1	1	15	-113,262652307749
392	2022-03-01	2022	3	1	3	1	2	1	16584,1341705322
423	2022-03-01	2022	3	1	3	1	2	2	1992,09766483307
454	2022-03-01	2022	3	1	3	1	2	3	102,699507236481
485	2022-03-01	2022	3	1	3	1	2	4	1164,61493378878
516	2022-03-01	2022	3	1	3	1	2	5	12,7836785763502
547	2022-03-01	2022	3	1	3	1	2	6	801,376574277878
578	2022-03-01	2022	3	1	3	1	2	7	339,361058473587
609	2022-03-01	2022	3	1	3	1	2	8	587,279444932938
640	2022-03-01	2022	3	1	3	1	2	9	1525,95473384857
671	2022-03-01	2022	3	1	3	1	2	11	195,567899465561
702	2022-03-01	2022	3	1	3	1	2	12	-13,0386326909065

Fig. 26. Tabela *Temp_DataSet_LinearRegression*.

5. Regressão de Ridge

O λ da função de custo da Regressão de Ridge regularizada foi definido como parâmetro de entrada da função *Python* que realiza o treinamento e a predição do modelo.

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \quad (18)$$

Após vários testes, foi definido $\lambda = 0{,}5$ e o resultado da previsão armazenado na tabela *Temp_DataSet_RidgeRegression* (ver Figura 27).

index	Data	Praca	Categoria	Tipo Veiculo	Forma Pagamento	Trafego
0	2022-03-01	1	1	Leve	pagante;previsao;	13392,995576031
1	2022-03-02	1	1	Leve	pagante;previsao;	10877,7316592211
2	2022-03-03	1	1	Leve	pagante;previsao;	10910,1836905399
3	2022-03-04	1	1	Leve	pagante;previsao;	10942,6357218587
4	2022-03-05	1	1	Leve	pagante;previsao;	10975,0877531775
5	2022-03-06	1	1	Leve	pagante;previsao;	10795,9151574431
6	2022-03-07	1	1	Leve	pagante;previsao;	10828,3671887621
7	2022-03-08	1	1	Leve	pagante;previsao;	10860,8192200809
8	2022-03-09	1	1	Leve	pagante;previsao;	10893,2712513997
9	2022-03-10	1	1	Leve	pagante;previsao;	10925,7232827183
10	2022-03-11	1	1	Leve	pagante;previsao;	10958,1753140371
11	2022-03-12	1	1	Leve	pagante;previsao;	10990,6273453559
12	2022-03-13	1	1	Leve	pagante;previsao;	10811,4547496217
13	2022-03-14	1	1	Leve	pagante;previsao;	10843,9067809405
14	2022-03-15	1	1	Leve	pagante;previsao;	10876,3588122593
15	2022-03-16	1	1	Leve	pagante;previsao;	10908,8108435781
16	2022-03-17	1	1	Leve	pagante;previsao;	10941,2628748969
17	2022-03-18	1	1	Leve	pagante;previsao;	10973,7149062156
18	2022-03-19	1	1	Leve	pagante;previsao;	11006,1669375343

Fig. 27. Tabela *Temp_DataSet_RidgeRegression*.

6. Regressão de LASSO

Da mesma forma que a Regressão LASSO, o λ foi passado como parâmetro da função de treinamento e cálculo da previsão.

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \quad (19)$$

O melhor valor encontrado utilizando para a previsão utiliza $\lambda = 0{,}5$. O conjunto resultado foi armazenado na tabela *Temp_DataSet_LassoRegression* (ver Figura 28).

index	Data	Praca	Categoria	Tipo Veiculo	Forma Pagamento	Trafego
0	2022-03-01	1	1	Leve	pagante;previsao;	13426,8337267042
1	2022-03-02	1	1	Leve	pagante;previsao;	10865,683502045
2	2022-03-03	1	1	Leve	pagante;previsao;	10890,6351267966
3	2022-03-04	1	1	Leve	pagante;previsao;	10915,5867515482
4	2022-03-05	1	1	Leve	pagante;previsao;	10940,5383762998
5	2022-03-06	1	1	Leve	pagante;previsao;	10795,5541812242
6	2022-03-07	1	1	Leve	pagante;previsao;	10820,5058059758
7	2022-03-08	1	1	Leve	pagante;previsao;	10845,4574307274
8	2022-03-09	1	1	Leve	pagante;previsao;	10870,409055479
9	2022-03-10	1	1	Leve	pagante;previsao;	10895,3606802305
10	2022-03-11	1	1	Leve	pagante;previsao;	10920,3123049822
11	2022-03-12	1	1	Leve	pagante;previsao;	10945,2639297338
12	2022-03-13	1	1	Leve	pagante;previsao;	10800,2797346581
13	2022-03-14	1	1	Leve	pagante;previsao;	10825,2313594096
14	2022-03-15	1	1	Leve	pagante;previsao;	10850,1829841613
15	2022-03-16	1	1	Leve	pagante;previsao;	10875,1346089129
16	2022-03-17	1	1	Leve	pagante;previsao;	10900,0862336644
17	2022-03-18	1	1	Leve	pagante;previsao;	10925,0378584161
18	2022-03-19	1	1	Leve	pagante;previsao;	10949,9894831677

Fig. 28. Tabela *Temp_DataSet_LassoRegression*.

E. Validação dos métodos

A função raiz quadrada do erro-médio foi escolhida para validar os métodos (*RMSE – root-mean-square error*). O valor de *RMSE* ficaram altos por causa da variável objetivo, tráfego, que também possui um alto valor. Sendo assim a função foi normalizada conforme Equação:

$$nRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{n}\left(y_i^{real} - y_i^{previsto}\right)}{n}}}{\left(Max(Trafego^{Atual}) - Min(Trafego^{Atual})\right)} \quad (20)$$

Os valores obtidos estão descritos na tabela abaixo. Sendo que o nRMSE foi calculado para cada dia, praça e categoria do mês de março, e o % Erro (21) representa o erro do valor total do tráfego em relação ao valor total real. Quanto menor os valores o método conseguiu prever o valor de tráfego de forma mais precisa para o mês de março.

$$\%Erro = \frac{\sum Valor^{Previsto}}{\sum Valor^{Real}} \quad (21)$$

As demais métricas Receita e Eixo Equivalente são calculadas baseadas no tráfego. Seguem as Tabelas 3 a 8 com os resultados de cada modelo para cada uma das métricas.

1. Modelo empírico

	$nRMSE$	% Erro
Tráfego	0,016781	-7,44
Receita	0,027806	-9,04
Eixo Equivalente	0,018798	-5,24

Tab. 3. Resultado do modelo empírico.

2. Médias

	$nRMSE$	% Erro
Tráfego	0,013296	2,01
Receita	0,033916	7,93
Eixo Equivalente	0,024426	7,45

Tab. 4. Resultado do modelo utilizando as médias.

3. Regressão linear

	nRMSE	% Erro
Tráfego	0,010636	-2,34
Receita	0,023276	-4,38
Eixo Equivalente	0,017741	-4,55

Tab. 5. Resultado do modelo utilizando regressão linear.

4. Regressão polinomial

	nRMSE	% Erro
Tráfego	0,011199	-5,16
Receita	0,018240	-1,61
Eixo Equivalente	0,014291	-1,65

Tab. 6. Resultado do modelo utilizando regressão polinomial.

5. Regressão de Ridge

	nRMSE	% Erro
Tráfego	0,010606	-2,13
Receita	0,023514	-4,57
Eixo Equivalente	0,017863	-4,72

Tab. 7. Resultado do modelo utilizando regressão de Ridge.

6. Regressão LASSO

	nRMSE	% Erro
Tráfego	0,010643	-2,40
Receita	0,024485	-5,63
Eixo Equivalente	0,018502	-5,71

Tab. 8. Resultado do modelo utilizando regressão polinomial.

Os modelos de regressão linear praticamente empataram na sua capacidade de prever o tráfego. A regressão polinomial teve um resultado inferior, demonstrando assim que melhor linha que representa o conjunto possuí relação linear, não curva. Baseado nesses resultados o modelo escolhido para compor a solução foi a Regressão de Ridge. Isso porque a regularização L2 trabalha melhor a provável relação de dependência entre as variáveis Feriado e Dia da semana.

F. Implementar

Método Preditivo:

O fluxo de dados do SSIS para calcular e carregar as tabelas mês atual e mês referência (ver Figura 29).

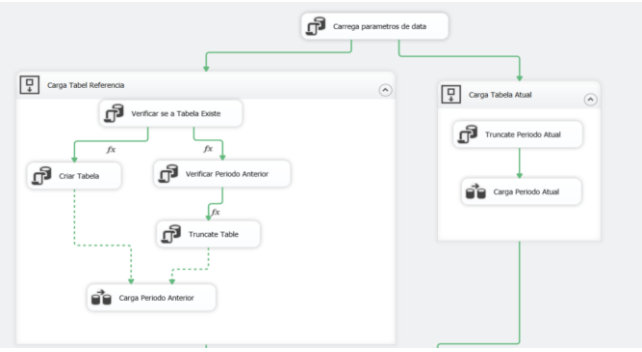

Fig. 29. Parte do fluxo de dados para escrita das tabelas com os valores de tráfego do mês atual e ano referência.

Primeiramente se calcula os parâmetros (ver Figura 30) de data para o mês que se pretende prever (geralmente o mês atual).

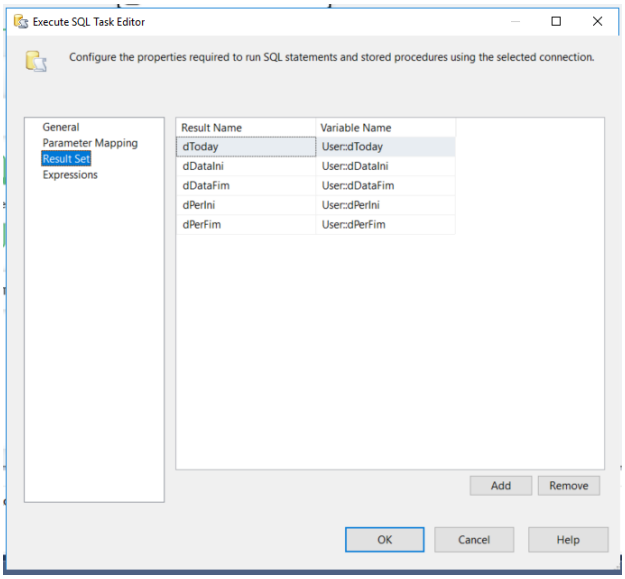

Fig. 30. Parâmetros de data para extração dos dados.

Posteriormente se extrai e carrega as tabelas para o armazém de dados. Em seguida é chamado um *script Python* que realiza as transformações, cálculos e cargas para criar a tabela com as datas previstas. Ele é baseado no método com melhor previsão (ver Figura 31).

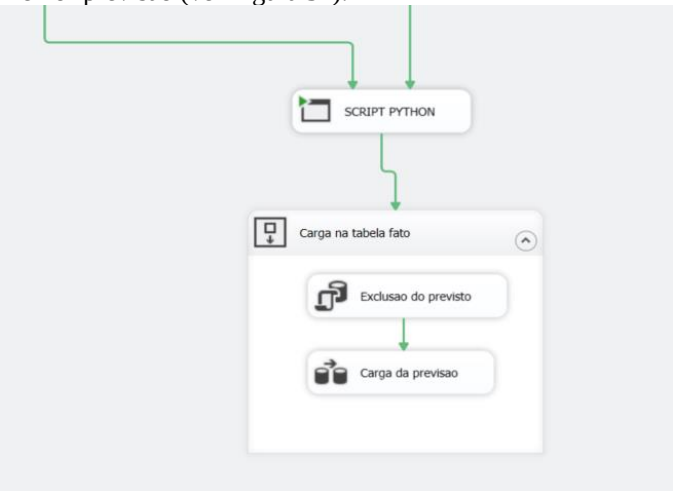

Fig. 31. Parte do fluxo de dados que chama o *script Python* e realiza a carga na tabela fato.

E como última etapa é agrupado as informações e inseridas na tabela fato.

Apresentação:

Para a apresentação e visualização dos dados foi utilizado o software Power BI. Para funcionar de forma mais eficiente foi escolhido o esquema estrela para o modelo tabular dos dados conforme imagem abaixo, no qual a tabela fato fica no centro e as tabelas de dimensões ficam ao redor (ver Figura 32).

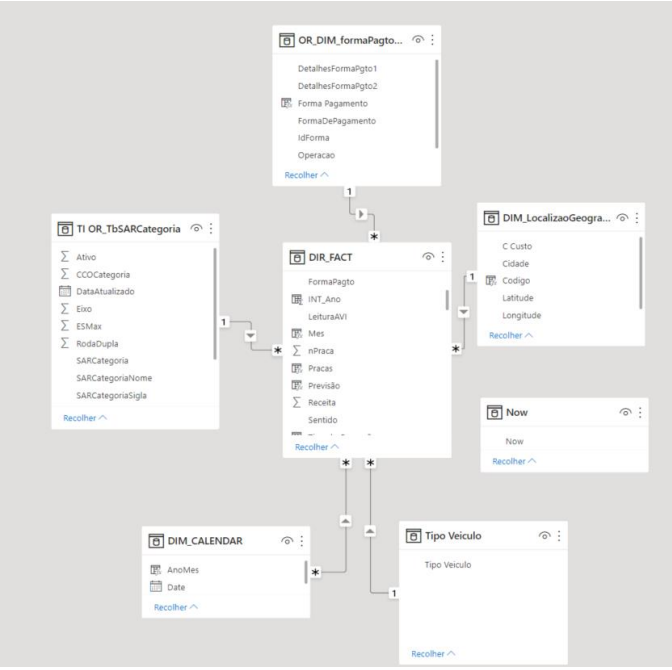

Fig. 32. Esquema estrela dos dados.

Com o objetivo de apresentar as métricas de forma inteligível e assertiva o relatório possuí nove páginas, cada uma delas apresenta as informações com dimensões e filtros diferentes, criando assim, diversos visuais possibilitando combinar e verificar o conhecimento obtido da melhor forma possível para os tomadores de decisão.

"Não se gerencia o que não se mede; não se mede o que não se define; não se define o que não se entende; não há sucesso no que não se gerencia" [11].

1. Comparativo Anual (Ano atual)

Essa página tem como objetivo apresentar de forma rápida e clara as comparações entre o ano atual e os anos anteriores, principalmente o último e o penúltimo ano. Foi inserido também a opção de se selecionar outros anos para realizar a comparação. São apresentadas duas métricas: o acumulado do ano e o acumulado do mês atual. Sendo assim possível comparar como está a performance da empresa em relação aos anos anteriores conforme Figura 33.

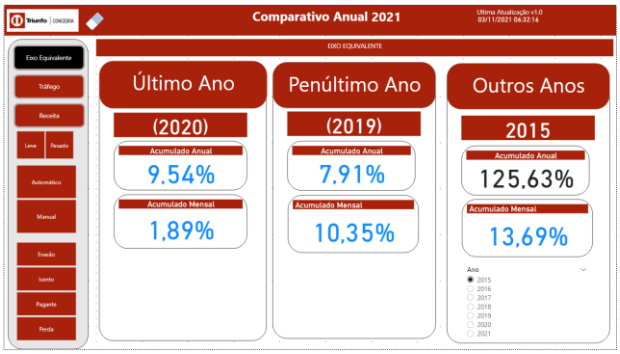

Fig. 33. Página Comparativo Anual Power BI.

2. Resumo Anual

Apresenta de forma resumida os dados do ano atual, permitindo visualizar como está o andamento da empresa no decorrer do ano (ver Figura 34).

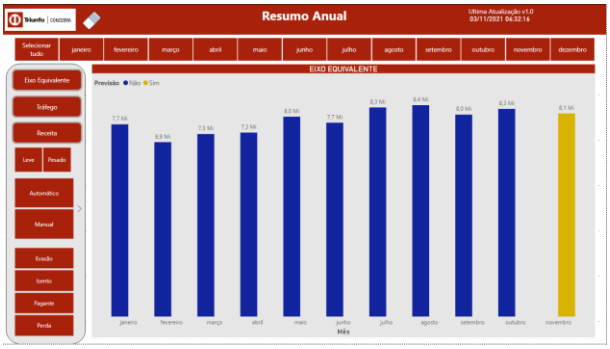

Fig. 34. Página Resumo anual do Power BI.

3. Evolução Anual

A página mostra a evolução da empresa no decorrer dos anos possibilitando a realização de *Drill Down* das informações. Com os diversos filtros disponíveis nessa página é possível realizar diversas análises (ver Figura 35).

Foi adicionado um *Tool Tip* (ver Figura 36) para quando passar o mouse em cima de algum visual mostrar para o usuário a representação dos valores para as dimensões Praça de pedágio e Categoria.

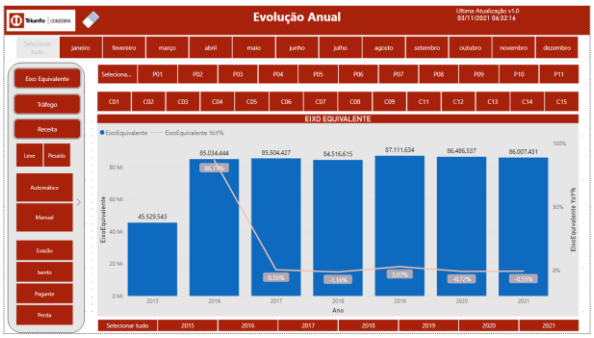

Fig. 35. Página Evolução Anual

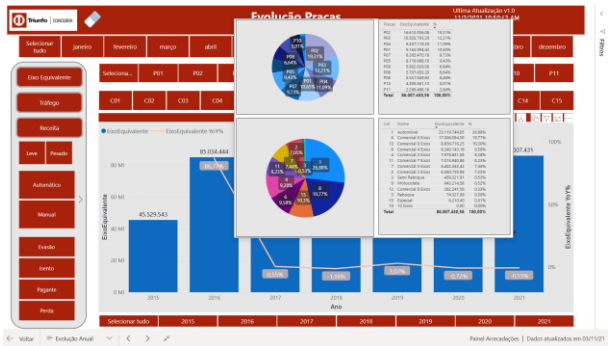

Fig. 36. Tool Tip com as informações de praças e categorias.

4. Evolução Anual das Praças de Pedágio

Com o intuito de facilitar as comparações entre as praças de

pedágio e como elas evoluíram no decorrer dos anos foi criado a página Evolução Anual das Praças (ver Figura 37), nela é possível realizar a análise do comportamento das praças de pedágio mês a mês. Adicionalmente há um *Tool Tip* (ver Figura 38) que permite visualizar os valores por forma de pagamento.

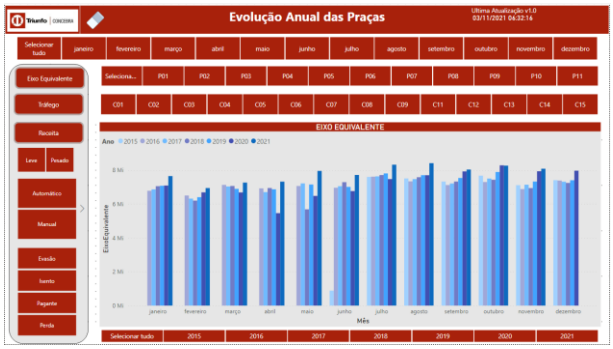

Fig. 37. Página Evolução Anual da Praças.

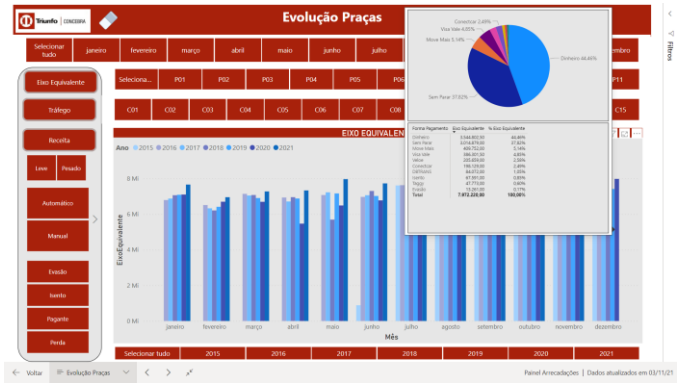

Fig. 38. Tool Tip para as formas de pagamento.

5. Mapa

Para se ter uma visão geográfica da concessionária foi criado um mapa, no qual é representado as praças de pedágio ao longo das rodovias da concessão. Através dele pode-se verificar a representação das praças de pedágio em relação as métricas e dimensões (ver Figura 39).

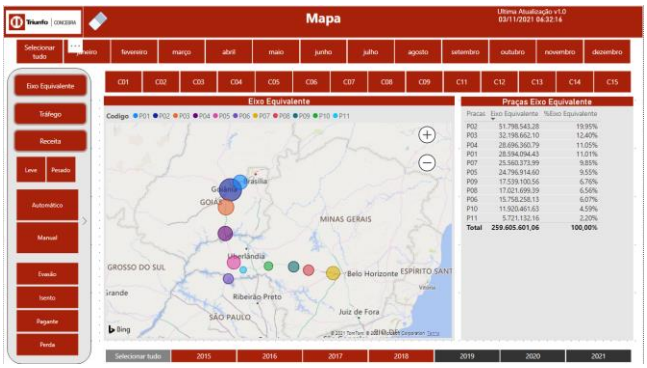

Fig. 39. Página com mapa com as informações das praças.

6. Avaliação dos Dados

A página Avaliação dos Dados foi criado com o intuito de se permitir uma avaliação mais bruta dos dados. Sendo assim, os dados foram apresentados em formato de matriz, possibilitando o *Drill Down* e filtro nas dimensões (ver Figura 40).

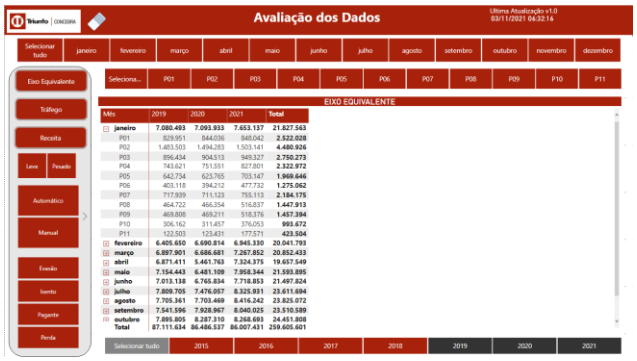

Fig. 40. Página com a matriz de dados brutos.

7. Evolução por Praça de Pedágio

Evolução por Praça de pedágio é uma página que mostra os dados de forma dinâmica (uma corrida de barras), possibilitando a visualização da evolução das praças de pedágio mês a mês (ver Figura 41).

Fig. 41. Página com corrida de barras mostrando a evolução das praças.

8. Evolução por Meio de Pagamento

Também um gráfico dinâmico (uma corrida de barras), a página Evolução por Meio de Pagamento demonstra a evolução mês a mês das formas de pagamento na concessão (ver Figura 42).

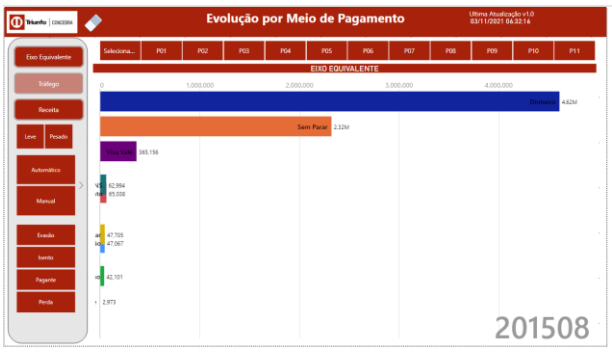

Fig. 42. Página com corrida de barras mostrando a evolução dos meios de

pagamento.

9. Evolução por Categoria

Evolução por Categoria, outra página que utiliza um gráfico dinâmico (uma corrida de barras) para demonstrar a evolução das categorias no decorrer da concessão (ver Figura 43).

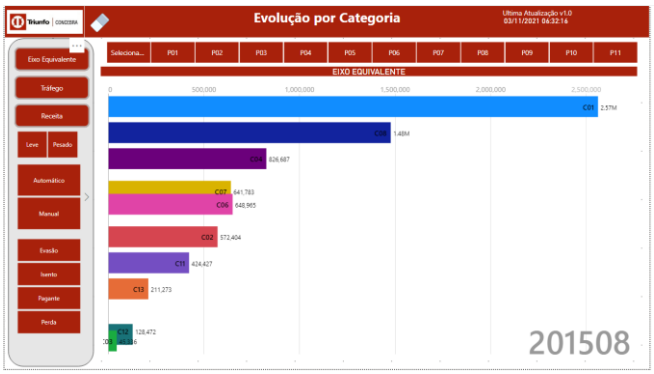

Fig. 43. Página com corrida de barras mostrando a evolução das categorias.

IV. Resultados e Conclusões

Mineração dos Dados

Os dados do problema inicialmente possuíam diversas fontes e um tamanho considerável. Com as técnicas de mineração foi possível unificar todas as fontes em um único banco de dados: o armazém de dados. Além disso foi possível reduzir significativamente o tamanho dos dados:

	Tamanho	Nº de Linhas
Base Sem Mineração	132,26 GB	289.137.813
Base Após Mineração	51,25 MB	349.784

Tab. 9. Tabela com os resultados da mineração.

Com a redução do tamanho a velocidade para acessar a informação também foi melhorada de uma média de 3 horas para 10 minutos.

Outro ganho foi o acesso à informação que é encontrada no portal do Power BI, e pode ser atualizada com um clique. Anteriormente levava um dia de trabalho e envolvia vários departamentos: Meio de Pagamento, Financeiro e Controladoria. Ocasionando erros na comunicação, anomalias e divergência nos dados.

Método Preditivo

Através dos resultados encontrados de *nRMSE* para a previsão do tráfego foi escolhido o modelo Regressão de Ridge como método preditivo.

Modelo	*nRMSE*	% Erro
Modelo empírico	0,016781	-7,44
Médias	0,013296	2,01
Regressão linear	0,010636	-2,34
Regressão polinomial	0,011199	-5,16
Regressão de Ridge	0,010606	-2,133
Regressão LASSO	0,010643	-2,349

Tab. 10. Tabela com os resultados da mineração.

Para verificar os resultados, foi feita uma comparação da previsão (15/03/2022) com os dados consolidados (04/04/2022), considerando as métricas de arrecadação.

	Tráfego	Receita (R$)	Eixo Equivalente
Previsto	3.338.616,95	22.101.047,66	7.455.311,31
Real	3.411.395,00	23.160.618,15	7.824.435,5
% Diferença	-2,13%	-4,57%	-4,72%

Tab. 11. Tabela com a comparação da previsão do mês de março de 2022.

Mediante a precisão encontrada, pode-se usar o valor previsto como base para análises, tomadas de decisão, construção de estratégias e táticas e visualizar objetivos.

As ferramentas de mineração de dados e um método preditivo acrescentam muito a capacidade de gestão e visualização de uma empresa. Possibilitando medir, visualizar e analisar a performance da arrecadação da concessionária.

Referências

[1] L. I. L. Faria and L. Quonian, Ferramentas para Estudos Prospectivos – Tutorial, 3º Workshop Brasileiro de Inteligência competitive e Gestão do Conhecimento, São Paulo, SP, Sep. 16-18 2002.

[2] J. Han, M. Kamber, “Introduction”, in Data mining: concepts, 2nd ed., Amsterdam, Elsevier, Netherlands, 2012, pp 1–42.

[3] U. Fayyad,G. Piatetsky-Shapiro, P. Smyth, “Form Data Mining to Knowledge Discovery in Data Bases”, AI Magazine, vol. 17, no. 3, pp 37–54, FALL,1996, DOI.

[4] D. Hand, H. Mannila, P. Smyth, “Measurement and Data” in Principles of Data Mining, 1rst ed., Cambridge, MA, MIT Press,2001, pp 21-37.

[5] S. Fortmann-Roe, “Understanding the Bias-Variance Tradeoff”, Fev. 25, 2022 [Online] Available: http://scott.fortmann-roe.com/docs/BiasVariance.html#fnref:1

[6] T. Hastie, “Overview of Supervised Learning” in The Elements of Statistical Learning Data Mining, Inference, and Prediction, 2nd ed, Stanford, CA, USA: Springer, 2008, ch. 2, sec. 9 , pp. 37 – 39.

[7] J. Shubham, “A comprehensive beginners guide for Linear, Ridge and Lasso Regression in Python and R”, Fev. 25, 2022 [Online] Available: https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/

[8] J. Hadamard, “Sur les Problèmes aux Dérivées Partielles et Leur Signification Physique”, Princeton University Bulletin, Princeton, NJ, USA, 1902, no. 13, pp. 49 – 52.

[9] L. J. Silvestre, “Regularização de Extreme Learning Machines: uma abordagem com matrizes de afinidade”, UFMG, Belo Horizonte, Brasil, Fev 2015.

[10] P-N. Tan, M. Steinbach, V. Kumar, “Data” in Introduction to Data Mining, 1st ed., Boston, MA, Pearson Education,2006, pp 19–88.

[11] W. E. Deming, “O método Deming de Administração”, 5a. ed., São Paulo: Marques Saraiva, 1989.

[12] P. M. Deshpande, S. Argarwal, J. F. Naughton, R. Ramakrishnan, “Computation of Multidimensional Aggregates”, Madision, WI, USA, May 1997.

[13] T. Anand and G. Kahn, “Opportunity Explorer: Navigating Large Databases Using Knowledge Discovery Templates”, Workshop Knowledge Discovery in Databases, Washington, DC, July 1993, pp 45–51.

[14] Y. Aumann, Y. Lindell, “A Statistical Theory for Quantitative Association Rules”, Journal of Intelligent Information Systems, December 2000. DOI 10.1145/312129.312243

[15] A. Ferrari and M. Russo, “Sharing the dashboard” in Introducing Microsoft Power BI,1st ed., Microsoft Press, Redmond, WA,2016, pp 27-47.
[16] A. Ferrari and M. Russo, “Understanding data refresh” in Introducing Microsoft Power BI,1st ed., Redmond, WA, USA, Microsoft Press, 2016, pp 49-61.
[17] A. Ferrari and M. Russo, “Using Power Bi Desktop” in Introducing Microsoft Power BI,1st ed., Microsoft Press, Redmond, WA,2016, pp 62-84.
[18] M. Russo and A. Ferrari, “DAX” in The Definitive Guide to DAX Business intelligence with Microsoft Excel, SQL Server Analysis Services, and Power BI, Microsoft Press, 1st ed., Redmond, WA, 2015, pp 1-129.
[19] C. N. Knaflic, “The importance of context” in Story Telling with Data, 1st ed., Hoboken, NJ, USA, John Wiley & Sons, 2015, pp 19-34.
[20] C. N. Knaflic, “Focus your audience’s attention” in Story Telling with Data, 1st ed., Hoboken, NJ, USA, John Wiley & Sons, 2015, pp 99-127.