UNIVERSIDADE CATÓLICA DO SALVADOR – UCSAL
MBA EM TECNOLOGIAS E APLICAÇÕES DE BUSINESS INTELIGENCE

TÉCNICAS E PROGRAMAÇÃO DE ANALYTICS – ANÁLISE PREDITIVA BÁSICA

Disciplina: Técnicas e Programação de Analytics.
Prof.: Luís Porciúncula
Aluno: Camila da Silva Oliveira

Salvador, Julho de 2019

Sumário

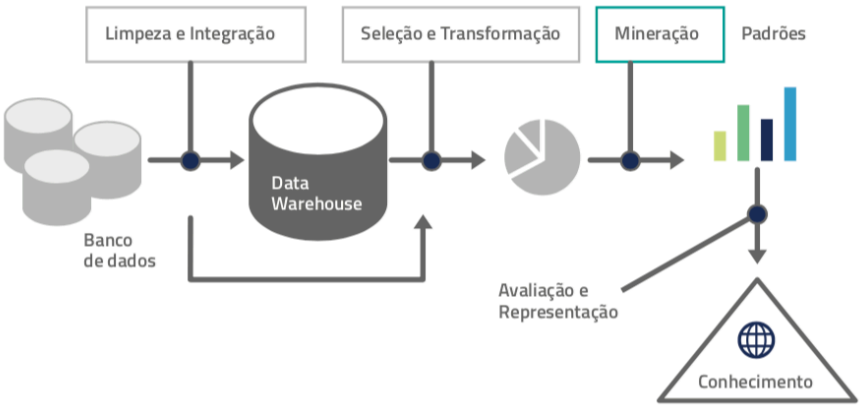
1. Introdução 3
2. Mineração de Dados 4
2.1. Entendimento dos dados disponibilizados 4
2.2. Limpeza e Integração 4
2.3. Seleção e Transformação 6
2.4. Mineração dos Dados 7
3. Conclusão 9
4. Bibliografia 10

1. Introdução

Uma seguradora de veículos deseja melhorar o posicionamento dos seus carros de apoio no período de chuvas dentro das regiões de abrangência de atendimento. Para isso, necessidade de uma estudo que informe as características do trânsito de cada bairro, quando chove e quando não chove. Como recomendação da empresa, recomenda-se incluir na análise o estudo por estação do ano.

Visando atender a necessidade da empresa, o estudo tem como objetivo apresentar o resultado da análise preditiva através da geração de uma árvore de decisão com base nos arquivos disponibilizados pela empresa, fornecendo insights para o problema apresentado. Para o desenvolvimento da atividade seguiu-se as etapas da metodologia KDD - Knowledge Discovery in Databases. O KDD é constituído pelas fases: limpeza e integração, seleção e transformação, mineração de dados e por fim avaliação dos resultados identificados. Para esta análise, foi inclusa uma etapa específica para entendimento dos dados disponibilizados.

Para o desenvolvimento do estudo, além do dicionário de dados fornecidos, foram utilizados os programas Excel, Google Maps e Studio R. Também utilizado para entendimento, validação da informação e consulta, o aplicativo Weka.



Fluxo mineração de dados – Fonte: BI como deve ser.

2. Mineração de Dados

2.1. Entendimento dos dados disponibilizados

No total foram disponibilizados cinco arquivos com informações variadas para possibilitar o desenvolvimento do modelo da árvore de decisão:

- Inmet_08_1963_03_2019.zip - Dados referentes à quantidade de chuva registrada em milímetros cúbicos no munícipio de Salvador entre agosto de 1963 a março de 2019. Observou-se que o dado disponibilizado não traz informações referente a localização da medição, impossibilita o cruzamento dos dados com demais fontes. Optou-se por não utilizar este arquivo, pois há um segundo com informações similar e com dados mais completo.

- Cemaden_ssa_01_2014_03_2019 – Dados referentes à quantidade de chuva registrada em milímetros cúbicos no munícipio de Salvador entre agosto de 1963 a março de 2019 nas estações de aferição do CEMADEN. O arquivo possui informações do índice pluviométrico, com detalhamento por estação de coleta, posicionamento geográfica data e hora da medição.

- Seg_log_bairro_regiao_ssa – informação da divisão dos bairros da cidade feita pela seguradora, a cidade é dividida em seis grandes regiões.

- Ra_bairros_pms – arquivo com a relação de bairros do município e a respectiva divisão administrativa.

- Seg_momentum – arquivo como o horário do trânsito é segmentado na cidade.

2.2. Limpeza e Integração

A etapa de limpeza e integração consiste no tratamento dos dados utilizados e montagem de uma base única e tratada para possibilitar as demais etapas da análise. O arquivo do Cemaned contém os principais dados deste estudo. Definiu-se o arquivo como base, sendo realizada junção com os outros dados para compor a base completa de análise. Todos os dados foram estudados e validados utilizando a ferramenta Excel.

Para o desenvolvimento da base do estudo, realizou-se: validação dos bairros, plotagem das estações de medição em mapa para identificação do bairro de localização de cada estação e levantamento das estações do ano para posterior cruzamento com o período de medição. Após limpeza foi realizada a composição da base única para possibilitar a análise e seleção do alvo do estudo.

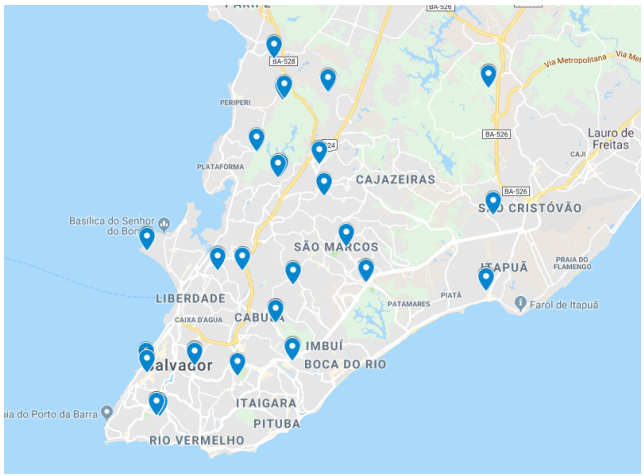
A validação do bairro foi efetuada através do cruzamento dos dados da base de segmentação por área com a relação de bairros da cidade de Salvador. Na relação disponibilizada pela seguradora não constam a informação de três bairros: Ilha dos Frades / Ilha de Santo Antônio, Ilha de Maré e Ilha de Bom Jesus dos Passos. Por se tratar de ilhas, onde não há pontes e o acesso consiste através de barcos e o número de carros na região é extremamente restrita, conclui-se que as informações da região estão corretas, podendo seguir com a análise.

Para melhor entendimento das informações e definição da estratégia de consolidação, efetuou-se a sumarização da estação por coordenada geográfica. Na primeira tentativa da plotagem da informação no Google Maps, verificou-se que alguns pontos geográficos estavam localizados no oceano, em outros o sistema sinalizou erro nas coordenadas fornecidas. Após depuração do erro, foi identificado que algumas informações de latitude e longitude estavam invertidas, outras estavam sem a separação do número decimal e outras sem dado. A correção foi efetuada caso a caso.

nomeEstacao	latitude	longitude
19 BC - Cabula	**#CAMPO!**	#CAMPO!
BATRE-São Cristovão	**-38368**	-12861
BNA-Paripe	**-38489**	-12794
EMBASA-Águas Claras	**-38441**	-12893
EMBASA-Alto do Peru	**-38485**	-12938
EMBASA-Alto do Pituaçu	**-38421**	-12943
EMBASA-Brotas	**-38495**	-12978
EMBASA-Gomeia	**-38474**	-12938
EMBASA-Pirajá	**-38458053**	-12898445
EMBASA-Rio Sena	**-38468**	-12888
EMBASA-Valéria	**-38437**	-12863
Hosp.Sarah-C.das Árvores	**-38453**	-12976
Inema-Monte Serrat	**-38516**	-12929
MAS-Dois de Julho	**-38515748**	-12981268
Politécnica - Federação	**-38511449**	-12999061

Tabela com informação de latitude e longitude inconsistente

Com as posições geográficas corretas, foi identificado o bairro em que cada estação de medição está situada ou a mais próxima possível, em alguns casos, o nome da estação coincide com o bairro, em outros casos foi efetuado a adequação pontual.



Google maps – localização das estações de medição.

Uma vez realizada a validação dos dados fontes, foi realizado o cruzamento dos dados e montado o arquivo único. Tendo como ponto de partida a base fornecida pelo Cemaned, conforme explicado anteriormente. Foi inclusa a informação do bairro onde está situada a estação de medição. Inclusão da informação da região, de acordo com a estrutura adotada pela seguradora, com base no bairro em que a estação está situada.

Inclusão da informação da estação do ano com base na data da medição e informação do momento do dia de acordo com o horário de medição. Uma vez estruturado e validado os dados em uma única base, definiu-se o escopo da análise.

Campo	Detalhamento Campo
codEstacao	Código da Estação de Medição
nomeEstacao	Nome da Estação de Medição
Bairro	Bairro da Estação
Area	Área da Seguradora
valorMedida	MM cúbico chuva
data_ssa	Data medição
hora_ssa	Hora de Medição
latitude	Latitude
longitude	Longitude
momento_transito	Dado do Trânsito
estação	Estação do Ano

Tabela com dados após limpeza e integração

2.3. Seleção e Transformação

O arquivo total se tornou extenso e pesado, sendo detectado alguns problemas de processamento da informação pela ferramenta devido ao tamanho dos dados. Optou-se por efetuar o estudo de uma das regiões centrais de posicionamento dos veículos pela seguradora, desta forma, será possível chegar a recomendações específicas, fornecendo melhores insights.

A região foco do estudo foi a de Brotas, pelo seu posicionamento e concentração de três bairros. A área de Brotas é composta pelos bairros de Brotas, Caminho das Árvores e Cosme de Farias. Serão analisados os dados de cinco estações de medição: EMBASA-Brotas, Caminho das Árvores, Hosp. Sarah – Caminho das Árvores, Brotas, Cosme de Farias.

Efetuado também a validação dos campos da tabela. Na realização das primeiras validações foram retiradas algumas colunas que não teriam relevância para a análise. Foram excluídas: o código da estação e o nome por serem fortemente relacionadas com a informação de bairro, criando redundância. A área por ser um valor único, latitude e longitude – a informação foi importante para a avaliação dos dados, mas não tem relevância na classificação da árvore de decisão.

Campo	Detalhamento Campo
Bairro	Bairro da Estação
valorMedida	MM cúbico chuva
data_ssa	Data medição
hora_ssa	Hora de Medição
transito	Dado do Trânsito
estação	Estação do Ano

Tabela com dados após seleção e transformação

2.4. Mineração dos Dados

No processo de análise dos dados optou-se por utilizar a linguagem R. Foi realizado uma série de teste para obtenção do melhor resultado. Na aplicação da técnica, efetuou-se uma segunda restrição dos campos utilizados e foram excluídos os campos de data e hora de medição. Após diversas aferições, chegou-se a seguinte parametrização do algoritmo utilizado:

- Fórmula genérica: trânsito em função dos campos bairro, estação do ano e o milímetro de chuva registrada.
- Método: Poisson, que expressa a probabilidade de uma série de eventos ocorrer num certo período de tempo.
- CP (parâmetro de complexidade): 0. Para possibilitar a exibição do resultado mais rico, optou-se por não otimizar o resultado da poda, sendo adotada restrição de valor mínimo para divisão.
- Minsplit: específico valor mínimo de 10.000 observações por nó para possibilitar a divisão.

```
library(readr)
library(rpart)
library(rpart.plot)
arquivo<-read.csv(file="/Users/camilaoliveira/Downloads/Brotas_py3.csv", sep=";", dec=",")
trans = rpart(transito~., data=arquivo, method="poisson", cp=0, minsplit=10000)
```

Código para análise definição da árvore de decisão.

Como resultado apresentado da árvore de decisão foi apresentado:

```
n= 113710

node), split, n, deviance, yval
      * denotes terminal node

  1) root 113710 125645.500 4.619576
    2) mm_chuva>=0.5 16958  17274.210 4.469397
      4) estacao=Outono,Verão 8853   8668.886 4.407438 *
      5) estacao=Inverno,Primavera 8105   8589.417 4.537078 *
    3) mm_chuva< 0.5 96752 108273.100 4.645899
      6) bairro=Brotas,Caminho das Árvores 79920  88997.760 4.634309
       12) estacao=Inverno,Outono,Verão 63664  70809.970 4.620885
         24) estacao=Inverno 24739  27719.570 4.608190
           48) bairro=Caminho das Árvores 14850  16643.940 4.606869 *
           49) bairro=Brotas 9889  11075.610 4.610173 *
         25) estacao=Outono,Verão 38925  43088.990 4.628953
           50) bairro=Caminho das Árvores 22016  24360.950 4.623047
            100) estacao=Outono 13770  15306.270 4.614379 *
            101) estacao=Verão 8246   9054.087 4.637521 *
           51) bairro=Brotas 16909  18727.650 4.636643
            102) estacao=Verão 7411   8111.005 4.625421 *
            103) estacao=Outono 9498  10616.290 4.645398 *
       13) estacao=Primavera 16256  18175.650 4.686884
         26) bairro=Caminho das Árvores 10032  11272.070 4.682315 *
         27) bairro=Brotas 6224   6903.462 4.694245 *
      7) bairro=Cosme de Farias 16832  19262.150 4.700926
       14) estacao=Inverno,Outono 8406   9600.747 4.695811 *
       15) estacao=Primavera,Verão 8426   9661.309 4.706027 *
```

Resultado aplicação *rpart* para árvore de decisão.

Ao plotar o resultado na imagem gráfica da árvore, obteve-se:

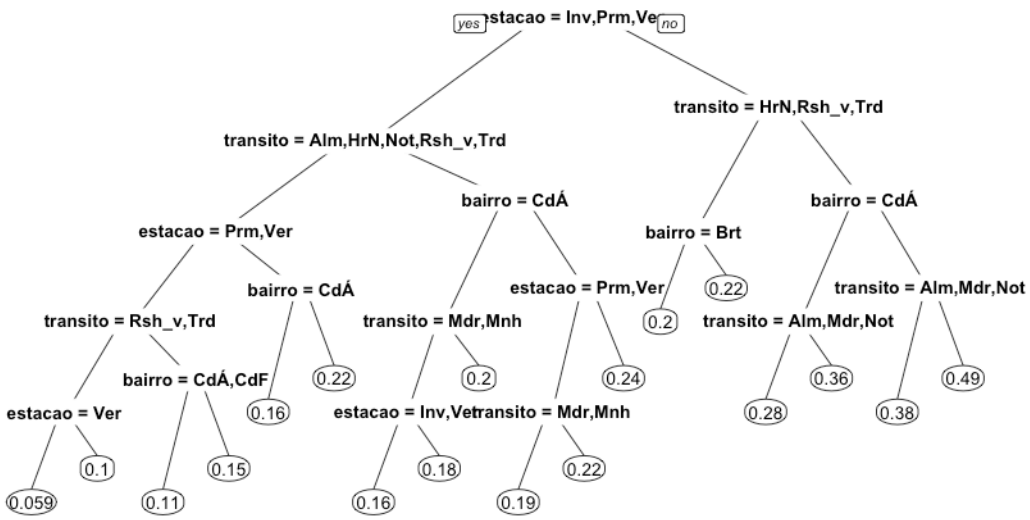


Imagem – resultado da árvore de decisão.

Como resultado da análise, observa-se que no outono há uma maior concentração de chuva. Durante esta estação, a incidência é maior no período entre a madrugada e no turno da manhã. Neste horário recomendasse concentração dos veículos entre os bairros de Brotas e Cosme de Farias, onde as chuvas são mais intensas.

Manter esta mesma localização no inverno, em especial no turno da manhã. As chuvas se intensificam no Caminho das Árvores no horário de rush matutino. A depender da estratégia da empresa, pode-se deslocar os veículos para esta posição. Nas demais estações a ocorrência de chuva diminui, mas há uma tendência de concentração também no bairro de Cosme de Farias e Brotas. Para definição mais complementa, recomenda-se complementar o estudo com o histórico de ocorrência.

3. Conclusão

O estudo apresentou uma excelente estratégia para definição do melhor local para alocação dos carros de suporte. Observa-se que uma estratégia que pode ser empregada é a permanência da base em locais distintos, a fim de melhorar o tempo de atendimento dos clientes da seguradora. Uma base flexível de acordo com o período do ano e combinado com uma análise histórica das ocorrências possibilitará reduzir o tempo de atendimento e aumentar a satisfação do cliente.

Alguns pontos de melhoria foram observados e recomenda-se aplicar no estudo para expansão do modelo para outras áreas. Foi adotada a premissa de vincular a localização da estação de medição ao bairro em que a mesma se encontra localizada, não abrangendo todos os bairros da região / cidade. Observa-se também que os dados do Cemaned possuem informações de coordenadas geográficas, pode-se criar uma nova árvore de decisão incluindo os dados de geolocalização. Desta forma, pode-se criar shapes com a região de abrangência dos postos de medição vinculados com a área de cobertura. Esta informação além de ser útil para estudos preditivos, também poderá ser utilizada na operação diária da seguradora.;

Complementar a isso, a inclusão do histórico de ocorrências registradas em cada bairro, enriquecerá a análise e possibilitará a criação de novos modelos preditivos, onde será possível trazer novos insights para o negócio.

4. Bibliografia

Oliveira, Diego Elias; Oliveira, Grimaldo Lopes de. BI Como Deve Ser – O Guia Definitivo. 2a ed. Salvador: 2016.

Distribuição de Poisson <https://pt.wikipedia.org/wiki/Distribuição_de_Poisson> Acessado em 28/07/19.