Arte dos Dados

Primeira parte do projeto prático

São José dos Campos - São Paulo
2024

Marcos Júnior Souza Gomes - marcosjsg1006@gmail.com

SUMÁRIO

1. SOBRE O PROJETO

Meu objetivo nesse projeto será desenvolver um mecanismo de precificação de seguros automobilísticos. Para o desenvolvimento do projeto, utilizarei os dados históricos da SUSEP (Superintendência de Seguros Privados) do ano de 2012 ao ano de 2021.

Para o modelo, esperasse que esse seja capaz de compreender as relações entre as variáveis que influenciam o procedimento de precificação do seguro para automóveis de passeio. Para além disso, é desejável entender quais as principais características analisadas no cálculo do prémio é mais influente em determinar os valores finais da apólice.

Com as atuais noções a respeito da precificação do seguro de um automóvel de passeio, esses são os objetivos atuais do projeto. À medida que mais ferramentas e conhecimento forem adquiridos ao longo do programa, espera-se que os objetivos finais do projeto sejam ampliados.

2. SOBRE OS DADOS

Os conjuntos de dados utilizados foram baixados através do site da SUSEP - Superintendência de Seguro Privado. No link, (https://www2.susep.gov.br/menuestatistica/Autoseg/principal.aspx), encontramos dados históricos de 2008 a 2021 sobre os seguros de diferentes tipos de automóveis.

Para todos os anos, os registros estão separados em três tabelas, sendo “arq_casco_comp”, “arq_casco_comp3” e “arq_casco_comp4”. Para a execução do projeto será utilizado apenas as informações descritas em “arq_casco_comp”. Esses bancos de dados estão organizados através das chaves Tarifária, Região, Modelo, Ano, Sexo, e Faixa Etária que são colunas no banco. Essas chaves são utilizadas para classificar e agrupar apólices que possuem essas informações iguais, ou seja, cada registro dentro desse banco representa um conjunto de riscos que compartilham as características das chaves em comum.

Para o banco “arq_casco_comp” de cada ano temos as seguintes colunas:

- **COD_TARIF** - representa o tipo de automóvel segurado pela apólice do conjunto de riscos compartilhados.
- **REGIAO** - região do território brasileiro onde o seguro foi contratado.
- **COD_MODELO** - código do modelo específico do automóvel segurado.
- **ANO_MODELO -** ano de fabricação do automóvel.
- **SEXO** - sexo do dono da apólice.
- **IDADE -** idade do usuário do seguro.
- **EXPOSICAO** - representa um estimador para a quantidade de veículos segurado dentro do grupo de risco compartilhado que foi definido.
- **PREMIO -** valor médio do prémio da apólice dentro do agrupamento de risco.
- **IS_MEDIA** - valor médio do limite a ser pago em um sinistro que contemple os riscos descritos pelo grupo.
- **FREQ_SIN1** - quantidade de sinistros da cobertura roubo\furto para o agrupamento.
- **INDENIZ1 -** total pago em indenização pelo sinistro de cobertura roubo\furto do conjunto de mesmos riscos.
- **FREQ_SIN2 -** quantidade de sinistros da cobertura colisão parcial.
- **INDENIZ**2 - total pago em indenização pelo sinistro da cobertura colisão parcial.
- **FREQ_SIN3 -** quantidade de sinistros da cobertura colisão perda total.

- **INDENIZ**3 - total pago em indenização pelo sinistro de cobertura colisão perda total.
- **FREQ_SIN4 -** quantidade de sinistros da cobertura de incêndio.
- **INDENIZ4** - total pago em indenização pelo sinistro da cobertura de incêndio.
- **FREQ_SIN9 -** quantidade de sinistros da cobertura de assistência 24 horas.
- **INDENIZ9** - total pago em indenização pelo sinistro de assistência 24 horas.
- **ENVIO -** data de envio do registro. Sendo XXXXA para o primeiro semestre do ano e XXXXB para o segundo semestre do ano.

Para cada ano são enviadas junto dos bancos principais algumas tabelas explicativas para os códigos utilizados em cada chave de classificação. Para o banco "arq_casco_comp" é de conhecimento que as chaves Tarifária, Região, Modelo, Ano, Sexo, e Faixa Etária estão descritas através de símbolos. Abaixo está a descrição das tabelas auxiliares que ajudam a compreender o que cada símbolo representa.

- **auto2_grupo** – código e descrição dos grupos de modelos.
- **auto_cat** – código de descrição de categorias tarifárias.
- **auto _ cau** – código e descrição de causas de sinistros.
- **auto_cob** – código e descrição de coberturas.
- **auto_idade** – código e descrição de faixas etárias.
- **auto_reg** – código e descrição de regiões de circulação.
- **auto_sexo** – código e descrição de sexo (masculino, feminino, jurídico).

3. CATEGORIA TARIFÁRIA

A princípio, adotarei para as análises futuras as categorias tarifárias de automóveis de passeio nacional, passeio importado, pickup (nacional e importado) e motocicletas (nacional e importadas) . Para o ano de 2012 ao primeiro semestre de 2021, no banco há 44.956.162 registros dessas 4 categorias tarifárias, sendo que isso representa 92,29% dos dados do arquivo "arq_casco_comp".

4. RISCOS NÃO ACEITÁVEIS

Como política de subscrição da empresa na categoria tarifária descrita, não serão cobertos danos ao automóvel decorrente dos seguintes riscos:

1) Uso do veículo para fins não declarados no momento da contratação da apólice. Ex: exposição a riscos decorrentes de atividade comerciais.
2) Avaria a casco em decorrência de vandalismo, desastre climáticos e guerra.
3) Dano de furto\roubo se averiguado pela investigação de sinistro que o proprietário do automóvel aceitou expor indevidamente o automóvel segurado.
4) Acidente em razão de embriaguez.
5) Sinistro causado por parente de primeiro grau ou cônjuge.
6) Mudança de proprietário do veículo durante a vigência do contrato, sem informar a esse segurador a respeito.
7) Omissão de informação durante a contratação da apólice.

5. SOBRE OS DADOS HISTÓRICOS

Como já mencionado, no site da SUSEP, onde se encontram os dados descritos, há registros disponíveis do ano de 2008 até o primeiro semestre de 2021. Para a execução do projeto que aqui está sendo descrito serão utilizados 8,5 anos de registro, que contemplam o período de 2012 a 2021. Tal época será adotada pois a partir de 2012 os dados da SUSEP começaram a ser disponibilizados em formato .csv que é mais fácil de ser manipulado. Os anos de 2008, 2009, 2010 e 2011 estão com a extensão .mbd ao qual não não se mostrou fácil de manipular.

6. RAZOABILIDADE

Após a junção dos bancos de tipo “arq_casco_comp” para todos os anos descritos nota se que apesar do grande volume de informação descritas uma parcela de apenas aproximadamente 5% do total dos dados possuíam suas linhas zeradas. Sendo assim, vemos que a grande maioria dos registros presentes estão em boas condições para a manipulação.

Anteriormente, visto que as variáveis EXPOSICAO, PREMIO, IS_MIN são todas descrições de outros conjuntos de dados que compõem os determinados grupos com risco compartilhados. Essas colunas do banco todas representam médias a respeito dos valores originais, sendo assim, podem estar sobre influência de outlier dos conjuntos originais. Dentro do banco faria-se necessário que cada uma das variáveis acima estivesse acompanhada do seu desvio padrão a fim de fornecer qual a dispersão dos dados em relação a média. Sem essa informação não será possível gerar um modelo em que possamos calcular a acurácia dos outputs para medí-las com precisão.

7. TRANSFORMAÇÕES PARA RAZOABILIDADE

A respeito do problema das médias descritas na seção anterior ainda não foi encontrado um método ou explicação para aumentar a razoabilidade dos dados categorizados em relação aos originais.

A respeito da consistência dos dados, para poder melhorá-la, as seguintes manipulações foram feitas no banco “arq casco comp”, utilizando-se a linguagem SQL.

1) Cada conjunto de dados baixado no site da SUSEP descreve um semestre de um ano. No total, para contemplar o período de 2012 a 2021 foi feito o download de 17 arquivos. Em todos esse foi necessário realizar a renomeação por semestre para facilitar a manipulação dos dados finais. Cada um dos arquivos “arq casco comp.csv” foram carregados e agrupados em um novo e único conjunto de dados denominado “arquivos_2012_2021”.

2) No dado original havia duas colunas denominadas “EXPOSICAO2” e “PREMIO2” que não possuíam nenhum registro, e portanto elas foram removidas.

3) Como já dito 5% dos registros estavam zerados em todas as colunas do banco, sendo assim, para evitar problemas nas análises todas essas linhas da tabela foram removidas.
4) Modificação das colunas para o tipo de dado correto. Foi necessário fazer um procedimento bastante longo para conseguir realizar essa tarefa, pois o código padrão em SQL para mudar o tipo de variável não funcionou.
5) Remoção dos registros de códigos tarifários que não serão utilizados na análise dos dados.

Infelizmente, não consegui fazer as demais modificações que haviam sido requeridas para esse módulo. O DataBricks começou a ficar muito lento a partir de certo ponto no código e demorava demais para processar essas etapas básicas. Como o cluster definido possui apenas 60 minutos, só para o carregamento e execução do notebook, estava demorando mais de 2 horas.

Tentei processar os dados diretamente no meu computador usando o python/jupyter, todavia, a RAM estava sendo insuficiente para armazenar todas as tabelas. Ainda, tentei fazer os mesmos procedimentos no Google Colab, entretanto há uma limitação muito grande de RAM na plataforma que não conseguir resolver.

Até a próxima etapa do projeto, espero achar um modo de contornar essas questões de processamento do banco.

8. SOBRE O TAMANHO DA AMOSTRA

O banco “arquivos_2012_2021” conta com 44.956.162 registros. Um número bastante alto de registro, o que é bom para fazer predições futuras. Todavia, é um grande desafio para o processamento de todos esses dados.

9. CREDIBILIDADE DOS DADOS

Para a realização do projeto futuro o tamanho da amostra parece ser bastante adequado, pois quando trabalharmos com os mecanismo de precificação através do Aprendizado de Máquina esse grande número de dados irá contribuir muito para aumento da acurácia dos resultados a serem obtidos.

Em relação a qualidade dos dados, em sua maioria, parecem ser bastante consistentes, mesmo para anos diferentes. A única ressalva a ser feita é a que tange ao problema da média acima descrito, que na melhor da circunstâncias será contornada posteriormente.

10. LIMITAÇÕES GERAIS

A limitação identificada foi descrita acima na seção 6.