Roberto Williams Batista
Udacity
Machine Learning Nanodegree
22 de abril de 2018
Versão: 3.0

Proposta de Projeto Final

Versão	Descrição
1.0	Versão inicial.
2.0	Alteração fundamental do projeto passando de análise de fraude para análise de gastos e predição de gastos.
2.1	Múltiplas correções.
2.2	Inserção de métrica R2.
3.0	Adição de feature na tabela 1.

PROPOSTA DE PROJETO

Histórico do assunto

Tendo em vista a repercussão de mal uso de cotas parlamentares[1] o momento político do Brasil surgiu a demanda maior transparência quanto ao gastos destas e outras verbas públicas. Um exemplo escandaloso do mal uso destas verbas é compra de reportagens por deputados[2] identificada pelo Projeto Marco Zero[3].

Com intuito de me engajar nesta realidade escolhi analizar o CEAP[4] e posteriormente publicá-lo nas mídias sociais para conscientização da população em geral, assim como incentivar outras iniciativas de monitoramento de gastos públicos.

Descrição do problema

A câmara dos deputados federais fornece dados históricos de gastos do CEAP organizados anualmente. Infelizmente não disponibilizam também um estudo de análise estatística destes gastos, assim como a previsão de comportamento de gastos do CEAP para o ano vigente baseado em dados históricos.

Conjuntos de dados e entradas

Serão utilizados os datasets dos gastos consolidados para os anos de 2015, 2016 e 2017, disponíveis no site abaixo:

https://dadosabertos.camara.leg.br/swagger/api.html

Os dados estão organizados de forma estruturado em três arquivos de formato CSV com 28 features (verificar tabela 1) :

A) Ano-2015.csv (376.862 linhas e 77.5MB)

B) Ano-2016.csv (358.238 linhas e 74.1MB)

C) Ano-2017.csv (341.222 linhas e 71.2MB)

Descrição da solução

A solução é composta pela aplicação de análise estatística descritiva e uso de algoritmo de regressão para predição de gastos para 2018. Serão utilizados as bibliotecas: Scikit Learn, Pandas, Numpy, Seaborn, OS,

Modelo de referência (benchmark)

Estatística descritiva

Serão utilizadas como referência as soluções disponíveis nos web sites abaixo:

a. **Precisamos falar sobre a Cota Parlamentar** - https://medium.com/data-science-brigade/precisamos-falar-sobre-a-cota-parlamentar-c58a73392148

   - Esta publicação realiza uma análise estatística dos gastos parlamentares e encontram relações interessantes entre gastos realizados. Esta é a principal referência a ser utilizada no projeto no tocante a estatística descritiva.

b. **Monitora, Brasil** - https://monitorabrasil.org

   - Este web site oferece a somatória de gastos dos deputados e também o ranking de maiores gastos dentre todos os deputados. É uma referência adicional ao projeto com alguns dados estatísticos descritivos.

c. **AKAN** - Acompanhamento de gastos de deputados - http://visualizemobile.github.io/akan/

   - Este aplicativo também pode ser utilizado como referência complementar no projeto.

d. **Operação Serenata de Amor** - https://serenata.ai

   - Este site tem referências estatísticas complementares e algumas abordagens que podem ser utilizadas neste projeto.

Predição de valor

e. Para as atividades relacionadas a regressão será usada o projeto de predição de valor de imóveis da cidade de Boston criado durante o curso e o projeto abaixo:

---

[1] https://medium.com/data-science-brigade/precisamos-falar-sobre-a-cota-parlamentar-c58a73392148

[2] https://theintercept.com/2018/01/16/deputados-usam-cota-parlamentar-para-comprar-reportagens/

[3] http://www.marcozero.info/

[4] Cota para exercício parlamentar.

https://www.coursera.org/learn/ml-foundations/lecture/2HrHv/learning-a-simple-regression-model-to-predict-house-prices-from-house-size

Métricas de avaliação

A solução poderá ser avaliada através do cumprimento da seguintes métricas:

a. Criação de gráficos de estatística descritiva e a análise de sua representação e potenciais descobertas.

b. Aplicação de algoritmo de aprendizagem supervisionada de regressão para predição de gastos para 2018 dos top 3 deputados de maior gastos em 2017 e análise de seus resultados. Será utilizado R2 Score (coefficient of determination) para determinação do score do algoritmo frente ao problema.

Design do projeto

De uma forma geral o projeto passará pelas seguintes fases:

1. Aquisição de dataset.
2. Exploração e ajustes iniciais dos três datasets.
    I. Importação de bibliotecas: Scikit Learn, Numpy, Pandas, Seaborn, OS, Matplotlib. Podendo ser adicionadas novas bibliotecas ou substituídas a listadas com objetivo de um melhor resultado do trabalho.
    II. Remoção de dados inválidos e uniformização de nomes e/ou termos utilizados.
    III. Confirmação dos features a serem desconsiderados na análise baseados na tabela 1 deste documento.
    IV. Criação de Pandas data frame com os dados a serem trabalhados.
    V. Pré-processamento de colunas não numéricas.
3. Análise inicial do dataset buscando gráficos e cálculos de valores absolutos e relativos utilizando-se de estatística descritiva.
4. Realizando somatória de gastos por tipo e por deputado, com resultado em um segundo data frame de trabalho. Este data frame será preparado para ser utilizado para aplicação do algoritmo de regressão. Como o objetivo inicial é utilizar a biblioteca Scikit Learn, devem ser analisadas os seguintes modelos de aprendizagem supervisionada disponíveis como potenciais algoritmos a serem testados. Seguem alguns exemplos de algoritmos potenciais:
    I. Support Vector Machine.
    II. K-Nearest Neighbors.
    III. Linear regression.
    IV. bayesian regression.
5. Separação de dados de treinamento, validação cruzada e teste para algoritmos de regressão.
6. Treinamento de algoritmos e realização de testes métricos para eleição de melhor algoritmo.
7. Análise do segundo data frame buscando gráficos e cálculos de valores absolutos e relativos utilizando-se de estatística descritiva.
8. Análise estatística sobre as previsões calculadas e observações.
9. Conclusões finais e exportação de arquivo CSV com a predição.

Tabela de features (tabela 1)

#	Uso	Elemento de Dado	Definição do Dado
1	S	**txNomeParlamentar**	Nome adotado pelo Parlamentar ao tomar posse do seu mandato.
2	N	**ideCadastro**	Número que identifica unicamente um deputado federal na CD.
3	N	**nuCarteiraParlamentar**	Documento usado para identificar um deputado federal na CD.
4	N	**nuLegislatura**	Legislatura: Período de quatro anos coincidente com o mandato parlamentar dos Deputados Federais.
5	S	**sgUF**	No contexto da cota CEAP, representa a unidade da federação pela qual o deputado foi eleito e é utilizada para definir o valor da cota a que o deputado tem.
6	S	**sgPartido**	O seu conteúdo representa a sigla de um partido. Definição de partido: é uma organização formada por pessoas com interesse ou ideologia comuns, que se associam com o fim de assumir o poder para implantar um programa de governo.
7	N	**codLegislatura**	Legislatura: Período de quatro anos coincidente com o mandato parlamentar dos Deputados Federais.
8	N	**numSubCota**	No contexto da Cota CEAP, o conteúdo deste dado representa o código do Tipo de Despesa referente à despesa realizada pelo deputado e comprovada por meio da emissão de um documento fiscal, a qual é debitada na cota do deputado.
9	S	**txtDescricao**	O seu conteúdo é a descrição do Tipo de Despesa relativo à despesa em questão.
10	N	**numEspecificacaoSubCota**	No contexto da Cota CEAP, há despesas cujo Tipo de Despesa necessita ter uma especificação mais detalhada (por exemplo, “Combustível”).
11	N	**txtDescricaoEspecificacao**	Representa a descrição especificação mais detalhada de um referido Tipo de Despesa.
12	N	**txtFornecedor**	O conteúdo deste dado representa o nome do fornecedor do produto ou serviço presente no documento fiscal
13	N	**txtCNPJCPF**	O conteúdo deste dado representa o CNPJ ou o CPF do emitente do documento fiscal, quando se tratar do uso da cota em razão do reembolso despesas comprovadas pela emissão de documentos fiscais.
14	N	**txtNumero**	O conteúdo deste dado representa o número de face do documento fiscal emitido ou o número do documento que deu causa à despesa debitada na cota do deputado.
15	N	**indTipoDocumento**	Este dado representa o tipo de documento do fiscal – 0 (Zero), para Nota Fiscal; 1 (um), para Recibo; e 2, para Despesa no Exterior.
16	N	**datEmissao**	O conteúdo deste dado é a data de emissão do documento fiscal ou a data do documento que tenha dado causa à despesa.
17	S	**vlrDocumento**	O seu conteúdo é o valor de face do documento fiscal ou o valor do documento que deu causa à despesa.
18	N	**vlrGlosa**	O seu conteúdo representa o valor da glosa do documento fiscal que incidirá sobre o Valor do Documento, ou o valor da glosa do documento que deu causa à despesa.
19	N	**vlrLiquido**	O seu conteúdo representa o valor líquido do documento fiscal ou do documento que deu causa à despesa e será calculado pela diferença entre o Valor do Documento e o Valor da Glosa.
20	S	**numMes**	O seu conteúdo representa o Mês da competência financeira do documento fiscal ou do documento que deu causa à despesa.
21	S	**numAno**	O seu conteúdo representa o Ano da competência financeira do documento fiscal ou do documento que deu causa à despesa.
22	N	**numParcela**	O seu conteúdo representa o número da parcela do documento fiscal. Ocorre quando o documento tem de ser reembolsado de forma parcelada.
23	N	**txtPassageiro**	O conteúdo deste dado representa o nome do passageiro, quando o documento que deu causa à despesa se tratar de emissão de bilhete aéreo.
24	N	**txtTrecho**	O conteúdo deste dado representa o trecho da viagem, quando o documento que deu causa à despesa se tratar de emissão de bilhete aéreo.
25	N	**numLote**	No contexto da Cota CEAP, o Número do Lote representa uma capa de lote que agrupa os documentos que serão entregues à Câmara para serem ressarcidos.
26	N	**numRessarcimento**	No contexto da Cota CEAP, o Número do Ressarcimento indica o ressarcimento do qual o documento fez parte por ocasião do processamento do seu reembolso.
27	N	**vlrRestituicao**	O seu conteúdo representa o valor restituído do documento fiscal que incidirá sobre o Valor do Documento.
28	N	**nuDeputadoId**	Número que identifica um Parlamentar ou Liderança na Transparência da Cota para Exercício da Atividade Parlamentar.