Proposta de Projeto Final

June 26, 2018

Lucas dos Santos Baião

1 Nanodegree Engenheiro de Machine Learning - Proposta de projeto final

1.1 Histórico do assunto

Desde de a metade do século XX a movimentação humana e os processos migratórios têm se intensificado sem precedentes, levando a uma grande quantidade de processos de expatriação e pedidos de vistos permanentes em diversos países. Notadamente, países com maior grau de desenvolvimento humano costumam ter uma alta taxa destas solicitações, e é cada vez mais comum a ocorrência deste tipo de situação a partir de propostas de trabalho em diferentes países. Em países com alta taxa de migração existe a tendência ao controle de entrada de estrangeiros com processos de pedido de visto.

Todo tipo de movimentação humana gera grande quantidade de informações, tanto para a população que migra, quanto para as localidades de origem e destino. A partir destes dados estatísticos é possível prever o sucesso, ou não, de uma solicitação de visto permanente. Para esta análise o país alvo sera os Estados Unidos da América, devido a sua grande taxa de crescimento populacional desde sua fundação (de aproximadamente 200 milhões para mais de 325 milhões de pessoas apenas no período entre 1975 e 2018, segundo Ediev [1].), sua grande taxa de imigração (mais de 96 milhões desde 1610, também segundo Ediev [1].) e o controle de fronteiras estabelecido de forma contínua e documentada.

Introduction

History of the US from the colonial times (from 1610) is marked by continuous population growth (fig.1) caused both by natural increase, immigration, and population aging.

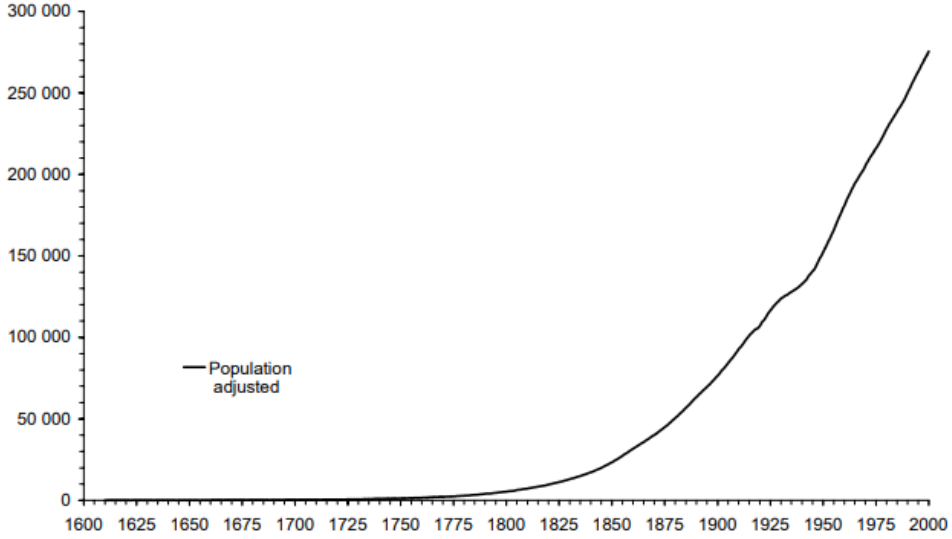

Figura 1. População dos EUA entre 1600 e 2000 [1]

Period	Immigrants settled
1610-1699	67 000
1700-1799	1 145 000
1800-1899	27 093 000
1900-1999	68 491 000
Total	96 795 000

Figura 2. Imigração para os EUA por período [1]

1. Reconstruction of the US immigration history:demographic potential approach. - Dalkhat M. Ediev (http://elibrary.lt/resursai/Uzsienio%20leidiniai/MFTI/2001/140e.pdf)

1.2 Descrição do problema

Pensando em auxiliar o processo de solicitação de vistos permanentes para os EUA e fazendo uso da grande quantidade de dados existentes sobre este tipo de processo, o intuito deste projeto é criar um sistema de classificação prevendo o sucesso ou não de uma solicitação para imigração para os EUA.

1.3 Conjuntos de dados e entradas

O conjunto de dados a ser utilizado foi obtido a partir da plataforma de datasets "Kaggle"[1]. Os dados estão em formato .csv, contendo 154 colunas (uma para cada parâmetro de solicitação) por 374.362 linhas (uma para cada processo de solicitação de visto). Entre estas colunas temos dados como: País de origem do solicitante, salário, cidade e estado de destino do solicitante, data, tipo de visto, além da informação de aceite ou não da solicitação.

As informações do dataset estão abertas a consulta na internet,tendo sido disponibilizadas pelo "Departament of Labour" (DOL) a partir de dados coletados pelo "Department of Homeland

Security's U.S. Citizenship and Immigration Services" (USCIS). Este Dataset conta com todos os registros de solicitação de visto permanente para os EUA no período compreendido entre 2011 e 2016, sendo aproximadamente 5 anos e meio de dados.

A maioria das colunas é de grande importância para a criação de um "perfil" do solicitante ao visto, sendo este primordial para a decisão das autoridades quanto a autorização ou não de imigração a um país, sendo este perfil uma forma de traçar a "probabilidade" de autorização ou não de entrada.

2. US Permanent Visa Applications: Detailed information on 374k visa decisions. (https://www.kaggle.com/jboysen/us-perm-visas)

1.4 Descrição da solução

Como proposta de solução para o problema apresentado acima, será desenvolvida uma solução baseada em machine learning, capaz de prever a autorização, ou não, de imigração para os EUA. O algoritmo será baseado na biblioteca Sci-kit learn, utilizando ferramentas baseadas em diversos tipos de algoritmo (regressões, SVM, redes neurais e etc...). A aplicação destas ferramentas se baseará na correlação entre os dados e os resultados de análises estatísticas sobre os mesmos.

A partir dos resultados dessa análise será possível identificar quais algoritmos se adequam melhor a solução esperada, de forma a gerar a classificação (visto concedido ou não) de forma assertiva.

1.5 Modelo de referência (benchmark)

O modelo de benchmark escolhido será "EDA US Permanent Visas with Feature Analysis"[1] de Bukun, publicado no próprio Kaggle. É um modelo desenvolvido na linguagem R, em que o Bukun realiza diversas features de exploração de dados, ótimas ferramentas de visualização (que tentarei implementar em python) e por fim desenvolve um algoritmo de XGBoost para realizar a mesma classificação a qual eu proponho.

O report de Bukun é o mais votado como "influente" para o dataset dentro do Kaggle e tem como resultado uma precisão de aproximadamente 71,85% utilizando cross validation. Minha proposta é usando os benchmarks e resultados alcançados por Bukun, desenvolver um projeto testando diversos tipos de algoritmos de Machine Learning diferentes, de forma a conseguir maior precisão nos dados de teste, além de ferramentas de visualização destes resultados, usando métodos gráficos para comparação.

O modelo de referência escolhido me ajudará, também, a perceber features abordadas pelo cientista de dados, que não conseguiria perceber sozinho em um conjunto de dados tão grande. A solução proposta por Bukun, embora bem trabalhada, apresenta apenas um tipo de algoritmo de machine learning implementado, que a pesar de bastante geral, pode não ser o melhor para a solução deste tipo de problemas.

1. (https://www.kaggle.com/ambarish/eda-us-permanent-visas-with-feature-analysis)

1.6 Métricas de avaliação

A principal métrica de avaliação será a acurácia do sistema (porcentagem de acertos das predições do algoritmo criado, utilizando como input uma parcela de dados de teste).

$$A_c = \frac{C}{p}$$

C = Acertos nas predições
P = quantidade de predições
Ac = Acurácia

Figura 3 - Equação da Acurácia de um modelo classificador

Levando em consideração o fato de que pretendo utilizar diversas técnicas de machine learning, além de feature engineering, cross validation e realizar fine tuning, utilizarei o melhor algoritmo em relação a acurácia e implementarei outras métricas a depender do tipo de algoritmo, em busca de eliminar riscos de underfitting, overfitting ou qualquer outro problema inerente ao tipo de algoritmo selecionado.

1.7 Design do projeto

O desenvolvimento do projeto se baseara ná criação de um algoritmo de classificação, a justificativa para esta escolha se deve principalmente a importância de delimitar apenas um "label" dentro dos campos do dataset [1], um indicador binário (sim ou não), a autorização ou não para o visto permanente nos EUA. A classificação poderá ser realizada por diversos tipos de algoritmo, como:

- GradientBoostingClassifier()
- DecisionTreeClassifier()
- RandomForestClassifier()
- LinearDiscriminantAnalysis()
- LogisticRegression()
- KNeighborsClassifier()
- GaussianNB()
- ExtraTreesClassifier()
- BaggingClassifier()

Este tipo de algoritmo deverá receber como input dados gerais do dataset e estimar, de diferentes formas, uma classificação para novos dados inseridos.

O fluxo de trabalho esperado será então:

Figura 4. Fluxo de Projeto

1. Import de bibliotecas e dados: Análise de quais bibliotecas, pacotes e tecnologias serão utilizadas, instalação e import destas ferramentas, leitura dos dados em formato csv e adequação

destes em um dataframe com sua correta identificação.

```
In [76]:
1  import pandas as pd
2  import matplotlib.pyplot as plt
3  import seaborn as sb
4  from plotly import __version__
5  from plotly.offline import download_plotlyjs, init_notebook_mode, plot, iplot
6  import cufflinks as cf
7  import numpy as np
8  init_notebook_mode(connected=True)
9  cf.go_offline()
10
11 df = pd.read_csv("us_perm_visas.csv", low_memory = False)
```

Figura 5. Código Inicial

2. Realização de data exploration e data visualization: Em busca de entender melhor os dados, seu comportamento e sua confiabilidade, desenvolverei códigos gerando gráficos, plotagens, tabelas e outros métodos de visualização, a fim de entender melhor como estes dados se organizam, quais são suas interdependências e correlações. Em conjunto com a visualização dos dados será realizada a filtragem de colunas dispensáveis, dados que não podem ser utilizados ou se encontram muito irregulares ou incompletos.

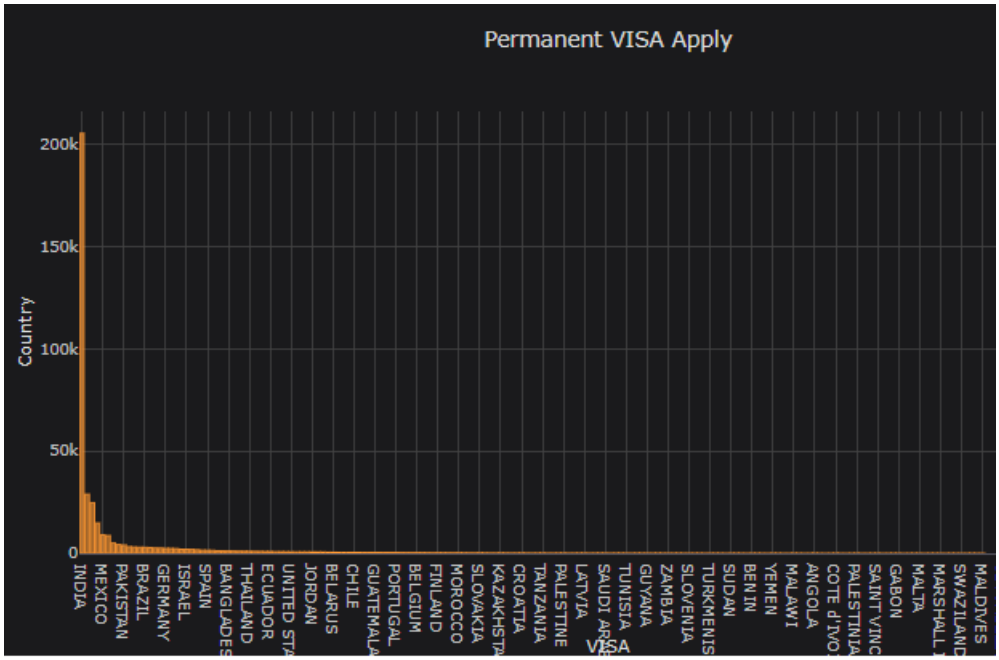

Figura 6. Exemplo de Data Visualization

3. Análise de Ferramentas: Busca por necessidade de utilização redução de dimensionalidade ou outro tipo de feature a ser utilizada para tratamento de dados ou feature engineering. Aplicação de ferramentas para melhoria ou adequação dos dados a utilizar nos classificadores.

4. Implementação dos classificadores: Serão realizados testes com pelo menos 3 tipos diferentes de classificadores em busca da melhor performance, serão implementados também métricas

como a descrição acima para verificação do melhor classificador.

5. Grid Search, Random Search e Fine tuning: O classificador com melhores resultados sera submetido a Grid Search e Random Search, de forma a obter os seus valores ótimos, utilizando diferentes métricas Pretendo ainda pesquisar outros métodos e parâmetros para realizar um fine tune final e obter os melhores resultados possíveis para a aplicação.

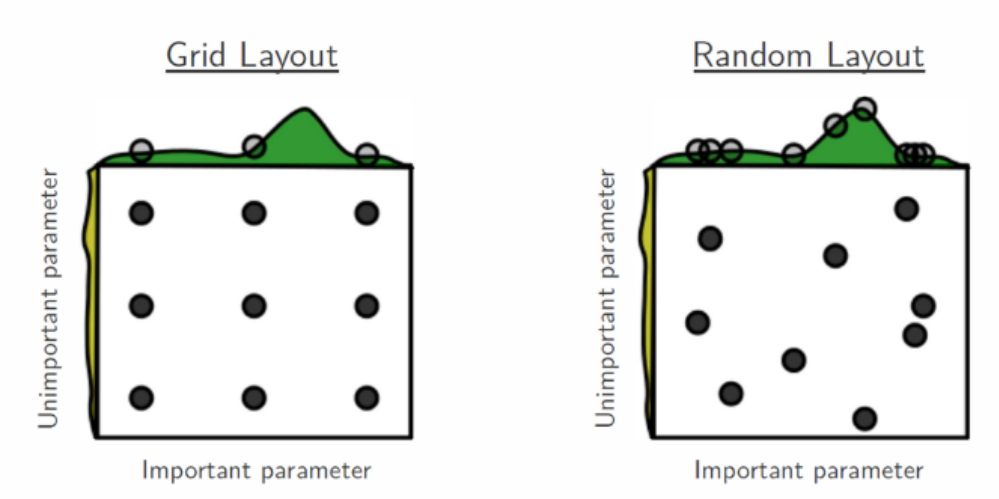

Figura 7. Grid Layout vs. Random Layout

6. Implementação do Classificador Final: Será implementado no código o classificador com melhores resultados de todos, serão avaliadas todas as métricas cabíveis a este tipo de classificador escolhido. Realizarei a plotagem de comparativos dessas métricas, além de outras demonstrações visuais de forma a facilitar o entendimento.

7. Testes Finais e Conclusão: Por fim demonstrarei alguns testes com perfis hipotéticos de forma a verificar seus resultados para aprovação ou não do visto, além de comentar as conclusões as quais cheguei ao fim da codificação.

Figura 8. Ciclo de Vida do Machine Learning em Produção [3]

1.(https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/)
2.(https://medium.com/rants-on-machine-learning/smarter-parameter-sweeps-or-why-grid-search-is-plain-stupid-c17d97a0e881)
3.(https://www.slideshare.net/turi-inc/model-management)