Taylor Imhof

Bellevue University | DSC 680

Portfolio Milestone 1 (Projects and README.md)

11/10/2022

List of portfolio projects

1. Python Car Price Prediction
    a. The primary purpose of this project is to train a predictive analytics model utlizing a used car dataset. The trained model would be provided to prospective car buying customers so that they would have a better understanding of the prices they should be paying based on the desired characteristics of the vehicle they want.
2. Reddit Headline Sentiment Analysis
    a. The primary purpose of this project is to conduct natural language processing (NLP), particularly sentiment analysis, on open text.
3. Flight Travel Safety Visualization Project
    a. The purpose of this project is to leverage the capabilities of Power BI to read in data and create useful data visualizations that can be used to convince an audience that traveling by air is safe.
4. OMDb Movie Data Preparation Project
    a. The first step of this project involved data collection. Data was collected using flat files .csvs, API calls, and databases. After the data was collected, the data was cleaned using the Python programming language. Once the data was cleaned, some data visulizations were created to get a peek at the underlying data trends.
5. Video Game Sales Exploratory Data Analysis

    a. The primary purpose of this project was to select a dataset of choice and perform exploratory data analysis (EDA) using the Python programming language.

6. Visualizing World Happiness with R
    a. The primary purpose of this project is to work with a raw dataset and create visualizations measuring world happines metrics gathered during the World Happiness Report

README.md Examples

Income Price Prediction Project

This project was worked on as part of a requirement for my Master's Data Science Program at Bellevue University.

-- Project Status: Completed

Project Intro/Objective

The purpose of this project was to work with a raw data set and peform machine learning to train a predictive analytics model.

Methods Used

* Machine Learning (Linear Regression, Logistic Regression)

* Data Visualization

* Summary Statistics

Technologies Used

* Python (Pandas, NumPy, Matplotlib, Statsmodels,etc)

* SQL

* Jupyter

Project Description

The first step in this project was collecting the data. The data was retrieved from Kaggle and imported into a Pandas dataframe. The dataset was a marketing dataset that contained a lot of interesting features such as income, education level, and the amount spend on different products. After the data was read in, there were some columns that needed to be cleaned. After the data munging stage, I used Matplotlib and Seaborn libraries to create visualizations to gain insights on underlying data trends. Once the data had been processed, two predictive models were trained. The first was a simple linear regression model using Statsmodel's OLS. The model was trained on the features and then used to predict the `income` column. The second model that was trained was a logisitic regression model. There were a few columns that measured whether a few targeted ad campaigns were successful. The logitic model was trained to predict whether the final ad campaign would be sucessful on future customers.

Reddit Headline Sentiment Analysis

This project was a requirement for one of my classes during my Master's Data Science program at Bellevue Univeristy.

-- Project Status: `In-Progress`

Project Intro/Objective

The primary purpose of this project is to conduct natural language processing (NLP), particularly sentiment analysis, on open text.

Methods Used

* Machine Learning (Sentiment Analysis: Vader, Roberta)

* Data Visualizaiton (Matplotlib, Seaborn)

* Summary Statistics

Technologies Used

* Python (Pandas, NumPy, nltk)

* DataSpell (Data Science IDE)

Project Description

The first step in this project was data collection. I leveraged the Python Reddit API Wrapper (PRAW) library to collect data from Reddit's exposed API they have for developers. I chose to collect information from [r/politcs](https://www.reddit.com/r/politics/). After the data was collected, I processed the data so that it could be stored in a Pandas dataframe. The primary Python package that was used for this project what the Natural Language Toolkit (NLTK) which comes packed with so many useful natural language processing capabilities. For this project, I only used the SentimentIntensityAnalyzer() function to genereate sentiment values in the form of polarity scores. These scores were combined into a useful `compound` metric that was used to encode a new label column that had distinct categorical values for `postiive`, `neutral`, and `negative` sentiment flavors. In order to analyze the performance of the trained sentiment analysis model, I used visualizations created using Seaborn and Matplotlib. I also created a WordCloud diagram that depicted the most common words measured in the dataset. For further detail on my project findings as well as the project methodology, please refer to the other files in this repo.