Instructions

The provided data represents restaurant inspections that took place in the Las Vegas metropolitan area. The original source of the data is located at the [City of Las Vegas Open Data Portal](https://opendataportal-lasvegas.opendata.arcgis.com/datasets/restaurant-inspections-open-data). Inspections are performed using a [Food Establishment Inspection Report](https://www.southernnevadahealthdistrict.org/download/eh/fe-inspection-report.pdf).

For this exercise, you will be given two data subsets: *TRAIN\_SET.csv* and *PRODUCTION\_SAMPLE.csv*.

**Provide a report explaining your responses to the tasks below**. There is no length requirement for this report - the amount of relevant insights you derive will determine its length. Your report should include your original code and be generated in HTML format via R Markdown, Jupyter notebooks, or similar.

While data scientists must demonstrate proficiency in EDA and model building, the **drift analysis task (3)** is a specific emphasis for the Model Steward role.

Tasks

1. **Conduct an exploratory data analysis (EDA) of the *TRAIN\_SET.CSV***. Provide an overview of the data and any underlying patterns you may identify. Without a thorough data dictionary, you may have to make some assumptions about the data. Document any transformations you perform. (*10 points*)

Work involving data transformations / wrangling is included and documented in the Python file.

EDA of the data was done in Tableau; sadly, I only have the trial version now with my student membership running out.

Chart, waterfall chart

Description automatically generated

Figure 1 Insights:

The largest amount of active Las Vegas Restaurant Inspection Grades come from Las Vegas. Following behind Las Vegas in terms of Total Current Grade Count is Henderson and North Las Vegas. I would assume that since Las Vegas is the most inspected city in this data set, it most likely also contains the most restaurants. It is also logical that the inspection distribution is skewed the way it is given Las Vegas’ population density and the fact that it is a popular tourism location

Chart, line chart

Description automatically generated

Figure 2 Insights:

The figure above shows the total number of inspections filtered down to Las Vegas over time. Also shown is the inspection type. It is evident that overall number of inspections seem to decline post-2012 for both inspection types. Concerning inspection types, most inspections seem to fall under the Routine Inspection, which is unsurprising. Re-Inspections post 2013 seem to stay constant in the [500,750] inspection count range.

Chart, line chart

Description automatically generated

Figure 3 Insights:

In the figure above, we observe a similar spike as we did for the inspections over time figure, with frequency of first inspections decreasing after 2012. It is important to note a slight increase in ‘Critical’ violation findings from 2013 to 2014, which is a reason why halfway through 2012 ‘Critical’ first violation findings were the most prevalent type.

Chart, line chart

Description automatically generated

Figure 4 Insights:

The figure demonstrates the same general pattern as Figure 3 with a few variations, one difference in the two figures being that in Figure 4 the ‘Major’ violation type does not exceed past the ‘Critical’ type. Additionally, the ‘Non-Major’ type is more prevalent in the beginning, but falls off, sometimes to 0, going into 2013.

Chart, line chart

Description automatically generated

Figure 5 Insights:

Violation Types other than ‘Non-Major’ seem to hold constant except for slight variations in the violation type ‘Major’ in years 2011-14.

Chart, waterfall chart

Description automatically generated

Figure 6 Insights:

This figure filters the data down to only the city of Las Vegas—the city with the highest number of inspections—to illustrate the prevalence of each type of restaurant. In Las Vegas, the most common types of restaurants are ‘Restaurant’ (meaning general restaurants), ‘Bar / Tavern’, ‘Snack Bar’, and ‘Special Kitchen’. There are more general restaurants than all other restaurant types combined, and more than 3 times as many general restaurants than there are bars or taverns, the second most common category.

Chart, bar chart

Description automatically generated

Figure 7 Insights:

The following figure shows the diversification in terms of total count of unique restaurant category’s each ZIPCODE is associated with, color-coded to show the distribution of cities within each ZIPCODE. The city of Las Vegas seems to offer the most variety in restaurant type. The cities of Mesquite, Henderson, Primm and North Las Vegas are also noteworthy with respect to containing a wide offering.

Chart, bar chart

Description automatically generated

Figure 8 Insights:

Nevada’s cities’ average inspection score is showcased in the figure above, with higher numbers of average inspection demerits being associated with worse grades. For example, the city with the best inspections scores on average seems to be Overton. The cities with poorer inspections, on average, include Boulder City, Goodsprings, Henderson, Las Vegas, Laughlin, Logandale, Mesquite, North Las Vegas, Primm, Sandy Valley, and Searchlight. There is enough claim to warrant a deeper analysis on whether distance from Las Vegas has a negative correlation with inspection demerits, on average.

A picture containing chart

Description automatically generated

Figure 9 Insights:

The figure above shows the count of all inspections that took place with a deeper look into the average inspection demerits for each inspection type while also offering insight on the grades given. Most inspections seem to fall into the Routine Inspection category and are graded ‘A’.

Chart, bar chart

Description automatically generated

Figure 9 Insights:

Above, the median employee age is shown graphed for each ZIPCODE in the filtered for cities. The median age group that most employees working in the restaurant industry fall into seems to be [18,34]. Based on our model, the median employee age of 34 is the max median across ZIPCODES, which I find interesting since I assumed there would be older individuals in the industry reflected in the dataset causing higher medians.

1. **Build a simple model that predicts the outcome of a restaurant’s next inspection, using NEXT\_INSPECTION\_GRADE\_C\_OR\_BELOW as the response**. Use your knowledge of the data and your own statistical expertise to develop an appropriate model. Document your thought process, the model techniques you considered, and the evaluation of the trained model. (*5 points*)

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, application

Description automatically generated with medium confidence**

**Table

Description automatically generated with medium confidence**

**Text

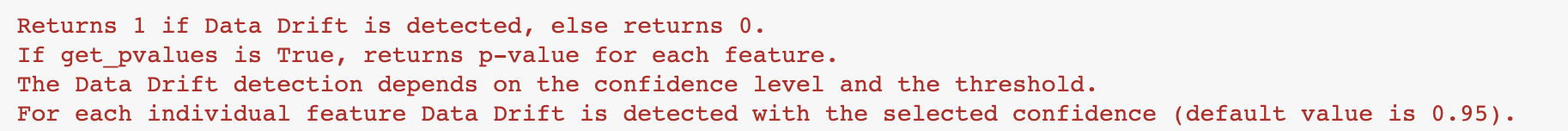
Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

1. Assume now that your model was deployed into production. The business partner is concerned because it appears to be underperforming on new observations. She has provided a production data extract for you to analyze: *PRODUCTION\_SAMPLE.csv*. **Perform a model drift analysis to help explain the performance discrepancy.** *Drift analysis identifies possible differences in the training and production datasets.* (*35 points*)

* What techniques or metrics did you use to identify potential drift?
* What columns have changed and in what ways?
* How would you assess the overall impact of the affected columns on the model performance?
* How would you address your findings with the business partner? What would you recommend?

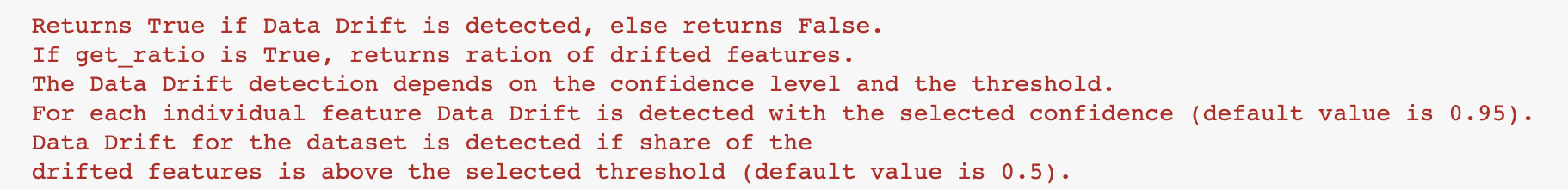


**FEATURE DRIFT**

1=RED (DRIFT) 0=BLUE (NO DRIFT)

Timeline

Description automatically generated



Timeline

Description automatically generated

**DATASET DRIFT** 1=RED (DRIFT) 0=BLUE (NO DRIFT)

Chart

Description automatically generated

Chart, bar chart, treemap chart

Description automatically generated

**How would you assess the overall impact of the affected columns on the model performance?**

* The historical drift analysis indicates that most features in our dataset are drifting over time, assuming our data collection methods remain constant model performance will continue to gradually worsen.

**How would you address your findings with the business partner? What would you recommend?**

* I would inform any business partners that our data is drifting. I would recommend looking into a larger sample of data over time to determine if the drift is not of interest i.e. in the case that the old model learned from historical data performs even better in classifications of current data stream.

Furthermore, I would advise looking into automating drift detectors in addition to automating refitting in the cases of drift. Additionally, we could weight the more drifting variables in real time to see changes in model performance.

1. **OPTIONAL.** How would you automate the model monitoring process in production to proactively detect drift?
   * What technologies/packages would you use?
     1. I would investigate using Evidently for the purpose of detection, but I imagine that if drift was detected we would need something in the data pipeline to address drift via weighting or expanding our data collection and our methods of collecting.
     2. I would also consider looking into how data bias may or may not cause drift.
   * What metrics would you track and how would you define appropriate thresholds?
     1. I would consider using any metrics Evidently has to offer at my current time. Creating a dashboard to track with would be another recommendation and is also something, depending on computational power, Evidently offers.
   * Where would the application it run?
     1. With the time given to complete this assessment I was able to detect drift using python. I would however imagine this application to be integrated within the data pipeline
   * How will the users consume it?
     1. Dashboard for visual consumption
   * Is there a monitoring job or is the monitoring a part of the model?
     1. For non-tech individuals a dashboard would be the best visual and would maybe be considered monitoring. The monitoring can be integrated somewhere in the data pipeline with the model.