San-Francisco Business Inspection: A Use Case of Food inspection by Scott W. Davis

In this report we will explore and analyze a dataset collected about San-Francisco businesses inspections. I have applied most of the stages of the data science methodology that I have studied in this specialization. This project will introduce a business inspection predictive analytics report that can help promote business safety and for example food business as part of the many processes put to prevent food-borne illness. Some of these processes include proper handling of food, proper preparation of food and its storage. Food inspection ensures that all these processes are done in such as a manner as to promote and achieve food safety.

Food inspection involves not only sampling and testing of end products but also assessing food centers to ensure compliance with food safety management systems. This minimizes the occurrence of public health food safety problems. Food inspection dates back to ancient times as part of the history of public health. The Food and Drug Administration (FDA) publishes the Food Code that sets guidelines and procedures to assist in food control jurisdictions. The Food Code provides a scientifically and legally backed basis for regulating the retail and food service industries. These include restaurants, grocery stores and institutional food service providers e.g. nursing homes. In the past, food inspection was done in a reactive manner whereby officers waited for reports of joints with possible non-compliance. However, it has been shown through research that food inspection should be done in a more proactive manner. Currently, some cities in the united states e.g. San Francisco are implementing a technologically driven approach to food inspection to try and predict food establishments that are more likely to be non compliant to food safety regulation. This is driven in part by the low Inspector to Food place ratio making it difficult to efficiently inspect all the food places. We will use Foursquare which comes with venue data that contains key descriptors of different venues including the category and popularity. This will show categories such as Nursing homes and food establishments along with attributes like name, address, ratings, and reviews from millions of points of interest. This report would be beneficial to public health specialists and every stakeholder working to alleviate public health concerns through preventive measures. The solution is not to introduce food inspection since these professionals are already carrying out food inspections in the relevant jurisdictions but to make the process more efficient.

In San-Francisco, it is estimated that one business inspector needs to efficiently inspect more than 500 business establishments given that there are only about 4 dozen inspectors to cover all business establishments. It is in waking of this statistic that the city saw an opportunity to make the process of food inspection more efficient by utilizing data analytics. In San-Francisco, through the Department of Public Health, systematically collected food inspection data from close to 100,000 sanitation inspections. Using this data, together with metadata on weather, related complaints e.g. sanitation, business characteristics, the city's advanced analytics team helped predict the food establishments that are more likely to violate food safety regulations. The food inspectors can then have a "Critical first" inspection

approach where the places that have been predicted to have critical violations are inspected first.

Some of the factors that tend to predict critical violation include previous critical violations, high temperatures, nearby sanitation complains, nearby burglaries etc

This report would be beneficial to public health specialists and every stakeholder working to alleviate public health concerns through preventive measures. It is not to introduce food inspection since these professionals are already carrying out food inspections in the relevant jurisdictions but to make the process more efficient.

Data Description

In this section I will the data that will be used to analyze the problem of food inspection and the source of the data. In order to develop a sufficient prediction system, the data should have the following categories:

Weather Data- In public health, the weather is a key component. Long rains are associated with flooding which predisposes to contamination of food with waterborne microbes.

- **Crime Data-** Higher crime rates have been strongly correlated with poverty due to lack of employment. Poverty has been in turn correlated with low hygiene which tends to predict the occurrence of critical violations of food safety regulations.

- **Places Data-** To help locate food establishments for inspection, there needs to be a way to pinpoint exactly where they are situated and preferably show it on a map. There are different sources of places data each which its set of strength and weakness.

- **Inspection Data-** Inspection data contains information such as previous the history of critical violations, type of facility, whether the establishment has a tobacco license and the length of time the establishment has been operating.

- **Water and Sanitation data-** Garbage and sanitation complaints can be used, together with other data, to try and predict critical violations. A place with frequent sanitation complaints is more likely to have a joint with critical violations as compared to another without any complaint.

- **Demographics data-** Demographics especially health demographics contain data about people living around a place including the age, sex, estimated income, occupation, recent infections all of which can be carefully correlated and used to predict a critical violation.

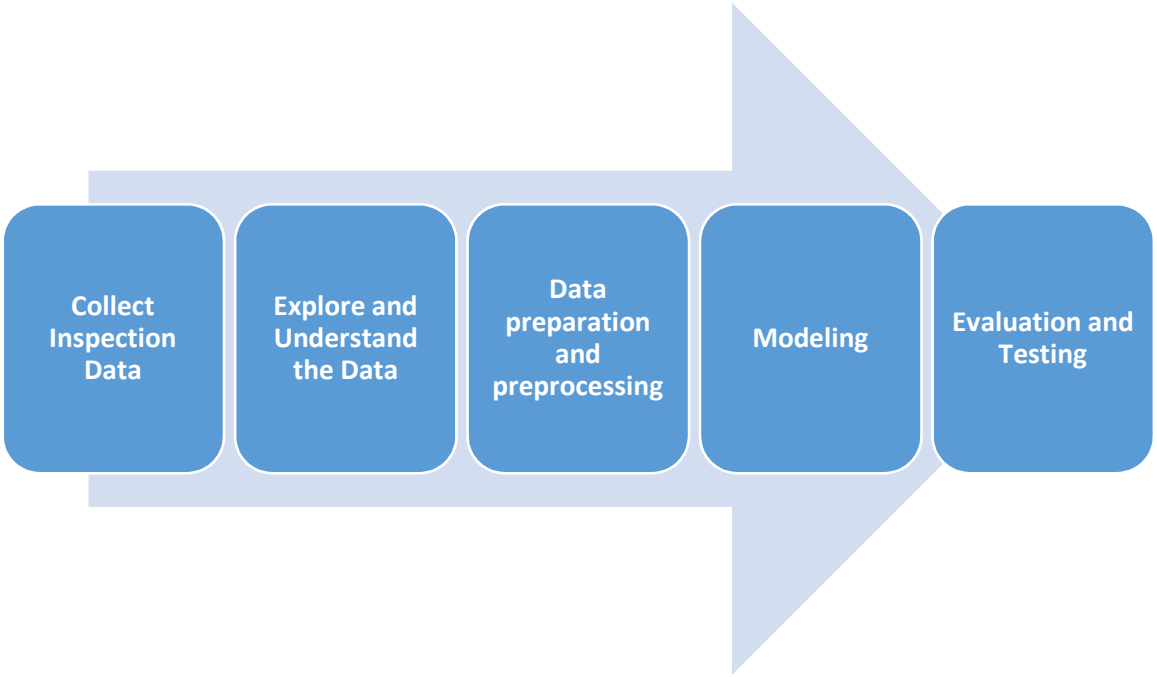
However, the data I have found is collected from (https://data.sfgov.org/Health-andSocialServices/Restaurant-Scores-LIVES-Standard/pyih-qa8i). The Health Department has developed an inspection report and scoring system. After conducting an inspection of the facility, the Health Inspector calculates a score based on the violations observed. Violations can fall into:

- **High risk category:** records specific violations that directly relate to the transmission of food borne illnesses, the adulteration of food products and the contamination of foodcontact surfaces.

- **Moderate risk category:** records specific violations that are of a moderate risk to the public health and safety.
- **Low risk category:** records violations that are low risk or have no immediate risk to the public health and safety. The score card that will be issued by the inspector is maintained at the food establishment and is available to the public in this dataset.

Methodology

In this part of the report we are going to describe the main components of our analysis and predication system. Our methodology consists of 5 components:



1. Collect Inspection Data

We downloaded the data from San-Francisco open data website as follows

```
url = "https://data.sfgov.org/resource/pyih-qa8i.csv"
wget.download(url)
sf_df = pd.read_csv('pyih-qa8i.csv')
```

```
5]: 'pyih-qa8i.csv'
```

The collected data are not ready for the analysis approach and need to be explored and organized.

2. Explore and Understand the Data

We read the dataset that we collect about San-Francisco business inspection into a pandas' data frame and display the first 5 rows of it as follows:

```
df_data_0 = pd.read_csv(body)
df_data_0.head()
```

Out[130]:

	business_id	business_name	business_address	business_city	business_state	business_postal_code	business_latitude	business_longitude	business_location	l
0	69618	Fancy Wheatfield Bakery	1362 Stockton St	San Francisco	CA	94133	NaN	NaN	NaN	
1	97975	BREADBELLY	1408 Clement St	San Francisco	CA	94118	NaN	NaN	NaN	
2	69487	Hakkasan San Francisco	1 Kearny St	San Francisco	CA	94108	NaN	NaN	NaN	
3	91044	Chopsticks Restaurant	4615 Mission St	San Francisco	CA	94112	NaN	NaN	NaN	
4	85987	Tselogs	552 Jones St	San Francisco	CA	94102	NaN	NaN	NaN	

5 rows × 23 columns

The dataset consists of more than 53k rows (inspection cases) and 17 columns (cases features or attributes). The following table give a brief description of each feature:

```
sf_df.dtypes
```

```
business_id                 int64
business_name              object
business_address           object
business_city              object
business_state             object
business_postal_code       object
business_latitude         float64
business_longitude        float64
business_location          object
business_phone_number     float64
inspection_id              object
inspection_date            object
inspection_score          float64
inspection_type            object
violation_id               object
violation_description      object
risk_category              object
:@computed_region_fyvs_ahh9    float64
:@computed_region_p5aj_wyqh    float64
:@computed_region_rxqg_mtj9    float64
:@computed_region_yftq_j783    float64
:@computed_region_bh8s_q3mv    float64
:@computed_region_ajp5_b2md    float64
dtype: object
```

We visualize the dataset to get more insight about it and discovering some pattern that might help in the modeling section. For more detail, we explain this section in details in ipython notebook, please check the file “Capstone Project - Final Project.ipynb”

3. Data Preparation and Preprocessing

In this component, we prepare the dataset for the modeling process where we choose the machine learning algorithms. To do that, we have cleaned the data from NaN values as follows:

```
copy_sf_df.dropna(subset=['business_id','business_name',
                          'business_address','business_city','business_state',
                          'business_postal_code','business_latitude','business_longitude',
                          'business_location','business_phone_number','inspection_id',
                          'inspection_id','inspection_date','inspection_score','inspection_type',
                          'violation_id','violation_description'],inplace=True)
```

We have extracted some new features from some fields. For example, from inspection_date we got the year, month and day and added them into the dataframe as follows:

```
copy_sf_df['year']= copy_sf_df['inspection_date'].apply(lambda x: getYear(str(x)))
copy_sf_df['Month']= copy_sf_df['inspection_date'].apply(lambda x: getMonth(str(x)))
copy_sf_df['day']= copy_sf_df['inspection_date'].apply(lambda x: getDay(str(x)))
copy_sf_df.head()
```

1]:

ıgitude	business_location	business_phone_number	...	risk_category	Neighborhoods (old)	Police Districts	Supervisor Districts	Fire Prevention Districts	Zip Codes	Analysis Neighborhoods	year	Month	day
419253	POINT (-122.419253 37.729016)	1.415546e+10	...	Moderate Risk	5.0	3.0	7.0	9.0	309.0	7.0	2019	09	03
419026	POINT (-122.419026 37.765142)	1.415563e+10	...	High Risk	19.0	4.0	7.0	8.0	28853.0	20.0	2019	07	24
405255	POINT (-122.405255 37.796152)	1.415599e+10	...	Low Risk	4.0	1.0	10.0	3.0	28857.0	6.0	2019	07	09
411684	POINT (-122.411684 37.776384)	1.415562e+10	...	Low Risk	34.0	2.0	9.0	8.0	28853.0	34.0	2018	11	27
418953	POINT (-122.418953 37.760295)	1.415578e+10	...	Moderate Risk	19.0	4.0	7.0	2.0	28859.0	20.0	2019	09	11

4. Modeling

After exploring the dataset and have a deep insight, we will apply a machine learning technique to classifying the inspection in order to have a better understanding of inspection process. We have three classes in this dataset as follows:

```
df_risk_score = copy_sf_df.groupby('risk_category', axis=0)['inspection_score'].mean()
df_risk_score.head()
```

]:
```
risk_category
High Risk        81.008163
Low Risk         87.345154
Moderate Risk    85.187436
Name: inspection_score, dtype: float64
```

To perform a machine-learning technique on this dataset, we have selected two main algorithms to do the classification:

- K Nearest Neighbor(KNN)
- Logistic Regression

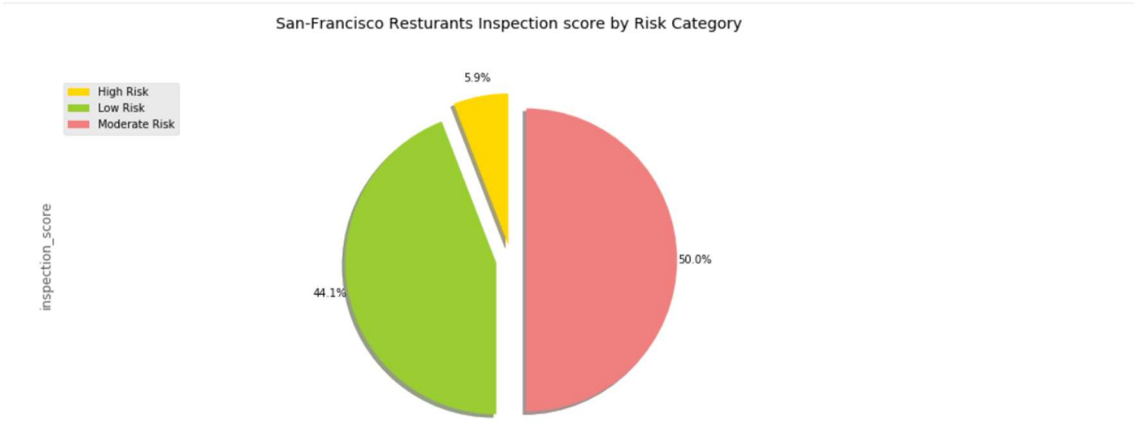
We have trained these two algorithms and tested them regarding our dataset.

5. Evaluation and Testing

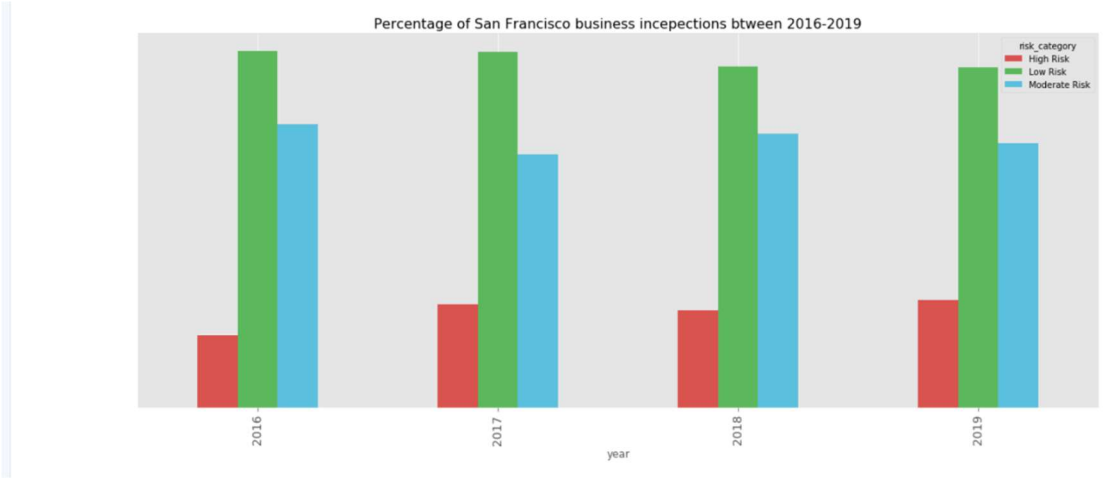
In this part we test the modeling algorithms by calculating the accuracy and f1-measure. We have also search for the best k that can give us the best classification model. In our case k was equal to 6.

Results

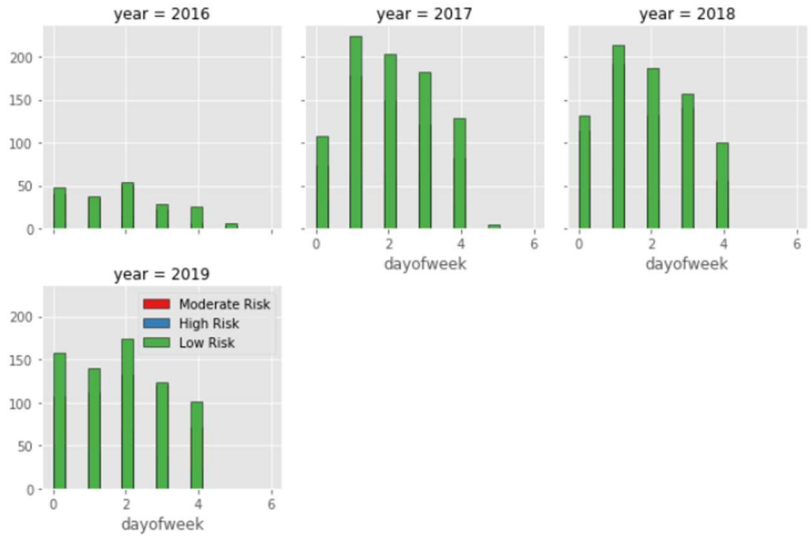
In this section, we can discus some results that we have got from the analysis and modelling sections. We have started by examining the categories of the inspections that we have in the dataset. We found that, in general, 44.10% of the businesses are considered in low risk, 50.00% are in moderate risk, while the high-risk businesses are 5.9% as depicts in figure bellow.



We grouped the inspections by year for each category low, moderate and high risk. We have found that the High Risk category increase from 2016 to 2017 more details are illustrated in figure bellow.



When we looked at the day of the week businesses were getting inspected, we have found that the inspection is very active in the beginning of the week and sharply decreased on Friday then increases little bit on Saturday as shown in figure bellow



Conclusion

To promote health, stakeholders in the healthcare industry need to continuously innovate to make the process more efficient. In food inspection, technology can be used to predict a likely critical violation through the use of data analytics instead of inspecting every joint blindly given the lack of enough manpower for this. The data used to predict critical violation include weather, crime and inspection data. Afterward, places data e.g. Foursquare is used to locate the food establishment for physical inspection.