Intro: 0.5p

* Business case
* Classification
* Enumerate methods

Data: 2p (2/3 per person)

What datasets

Data preparation

Analysis: 3p (1 per person

If time: discussion

Conclusion 0,5p

# Intro:

Meat, poultry, eggs -> unresponsible handling -> health risk (despite oversight: E. Coli, Salmonella, Listeria outbreaks hospitalized 131 and killed 6 people in the usa p.23)-> regulations and tight oversight required

USA -> specific agency: Food Safety and Inspection Service <https://www.usa.gov/federal-agencies/food-safety-and-inspection-service> (other food related agencies Food and Agriculture, National Institute of (NIFA)

Food and Drug Administration (FDA)

Food and Nutrition Service)

Food Safety and Inspection Service: Budget total: 1,03 – 1,05 billion US$, new York: 13,3 – 15,3 million US$ p.16

While the official budget report unfortunately does not include a breakdown of the costs directly caused by the inspections, the classification by objects shows that the vast majority of funds, a range of 960 – 990 million US$ (2016 – 2019), are connected with inspections, suggestions them to be among the most central cost points of the organization. P.17

Means: if we could predict which food retailers pose the highest public health risk and allocate resources efficiently, this could not only lead to cost reductions of millions of US$ every year but might even save lives by preventing the outbreak of dangerous foodborne illnesses.

Such a predictive algorithm could be based on the already existing data of the grades (scale: A (no Issues) over B (Minor Issues) to C (major Issues)) received by food retailers during past inspections, and variety of additional variables, including but not limited to customer experiences (based on google ratings) and demographic information.

In the following section we set out to do exactly that: Working with the data provided by past inspections matched with various additional variables, we experimented with various predictive algorithms to discover a way of estimating the future grade of a food retailer.

# Demographic data

Since the original dataset did not feature any demographic data, related to the inspected stores, this type of data had to be acquired otherwise and matched subsequently to the inspection dataset. The obvious choice for the matching was to match by location. Fortunately, the amount of demographic data on different sections of the state New York is ample, in the form of different datasets detailing the results of various census-endeavours. Unfortunately, a large percentage of these datasets turned out to be unusable. This dataset, for example, includes demographic data by zip code, which would have provided a straightforward way of matching with the inspection’s dataset, turned out to be based on sample size oftentimes just one or even zero questioned persons. We therefore chose to use two datasets based on US wide census studies, by the official census bureau of the USA. This information can be downloaded from the American FactFinder <https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml>

But are also available on Kaggle

<https://www.kaggle.com/muonneutrino/us-census-demographic-data>

which we ultimately used. The first dataset provided us with demographic information, such as employment status and ethnicity, by US-County. The datasets could be very easily matched via the variable county, but due to the large size of the counties, the information did not necessarily exactly reflect the demographic makeup in the direct vicinity of the food retailers. The second dataset, which displayed the demographic makeup of the US, by Census Tract, the relatively small geographic locations specifically designed for the collection of census-data, offered notably more specific information about the retailers’ surroundings.

Since Census Tracts, however, do not follow the traditional pattern location specification used in the inspection’s dataset, direct matching of the two set was not possible. An approach of nevertheless matching the sets, was found in the creation of a third “translation” dataset AddTrac (short for Address to Census Tract) which matched the retailers’ locations with the Census Tract codes. To do so, we made use of the census bureau’s geocoding service, which can be found here

<https://geocoding.geo.census.gov/geocoder/geographies/addressbatch?form>

This service added various geolocation-identifiers to specifically created csv files holding the addresses of our food retailer. Unfortunately, circa 20% of the addresses were not be identified by the service (possible reasons include name changes and confidentiality concerns <https://www2.census.gov/geo/pdfs/maps-ata/data/FAQ_for_Census_Bureau_Public_Geocoder.pdf>, p. 7

which led to some data loss. By combining the now known state, county FIPS codes and census block codes, as well as adding placeholder 0 where necessary, the Tract Ids could be recreated. Having the Tract Ids, it was possible to merge the more exact Census Tract based data with the inspections dataset.

# Conclusion

Despite the utilization of an array of different variables and methods, the prediction of Inspection Grades proved to be exceedingly difficult.

#short recap by method of how well it went

It seems likely that the reason for this regrettable result lies in the low predictive quality of the here used independent variables. We have already showcased the, if anything, very weak correlations between the data that had been available to us and the dependent variable. It seems that the stores inspection grade is, if at all, only very weakly dependent on the customer experience as shown by google-ratings. A close look at deficiency descriptions illustrates why this might be the case: The deficiencies might be very minor or unnoticeable in the eye of the customer; the store does not have to be completely filthy to fail an inspection.