What makes you ill?

IST 718 | Advanced Information Analytics | Anmol Handa | Anuj Jain | Justin Thierry

Problem

One in every person in the United States gets sick from eating food at various location in different states. While most foodborne illnesses are not considered outbreaks. But if some of them are at a larger scale, they can be life threatening. The problem at hand is the Foodborne outbreaks in the past caused by various factors. We want to find a way to predict the major factors that contribute towards foodborne diseases.

Data Description

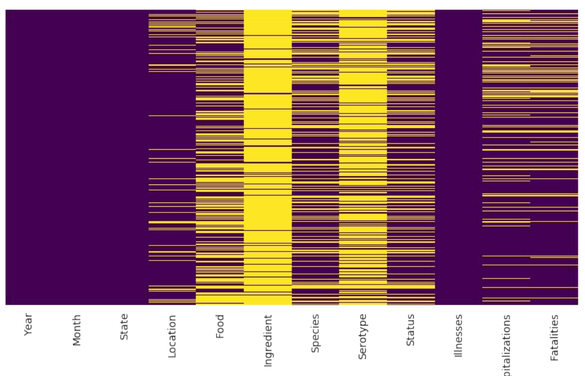
The dataset consists of 12 rows and has about 19,000 records. It consists of data ranging from 1998 to 2015 and across various states, months and location where the food was prepared. It contains the type of food column that causes illness. Other columns are sparsely filled and roam around pathogens and ingredients of food.

Objectives

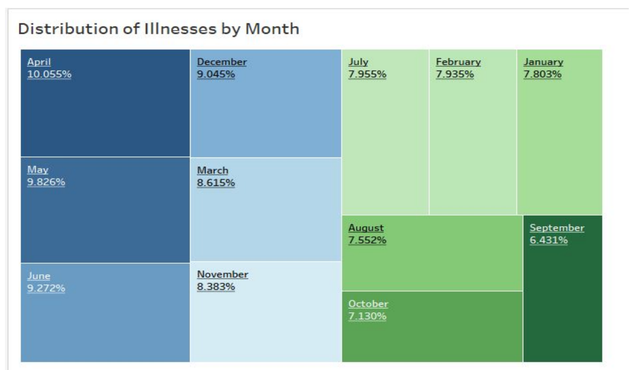
- Exploratory Data Analysis to analyse the dataset for summarising the main characteristics and distribution of illnesses
- Performing trend analysis to comprehend the variance of illnesses with respect to time
- Identifying the major features causing the outbreaks
- Applying Regression modelling techniques like Linear Regression, Decision Tree Regression and Random Forest Regression to minimize the root mean square values while predicting the illnesses on a normalised scale
- Applying Logistic Regression modelling to understand the magnitude of illnesses as high or low based on area under the curve and confusion matrix

Data Exploration graphs

Null values identification

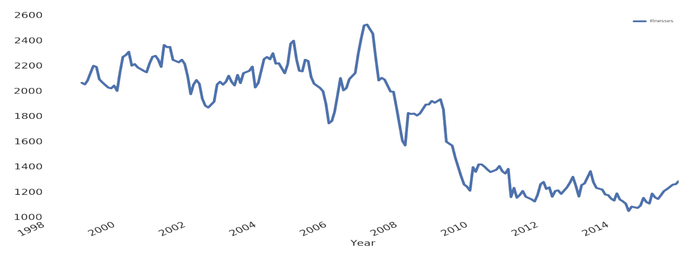


Distribution of Illnesses by Month

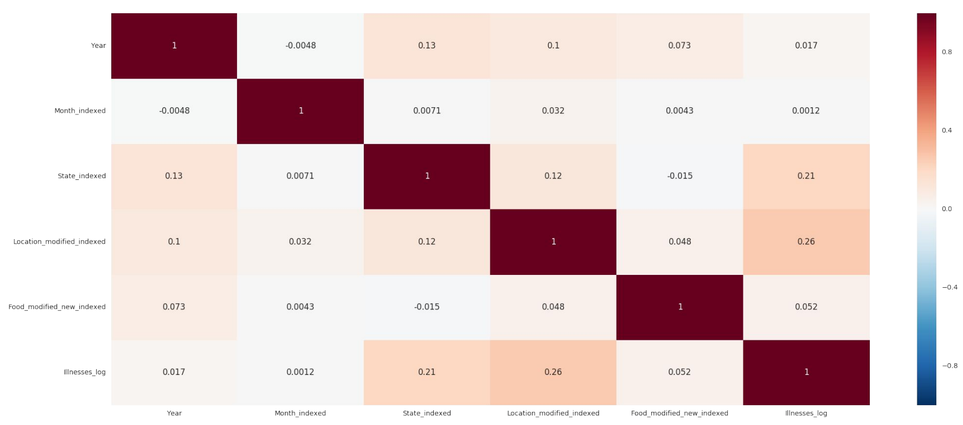


Distribution of Illnesses by State

Time series analysis for Illnesses

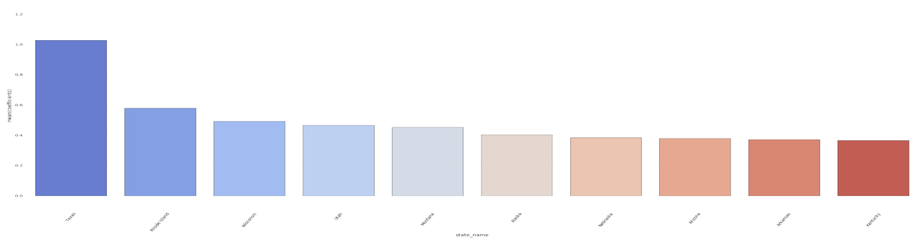


Correlation between major features

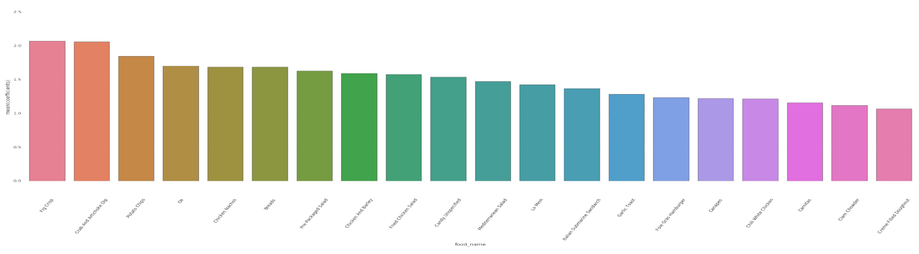


Feature Importance Graphs

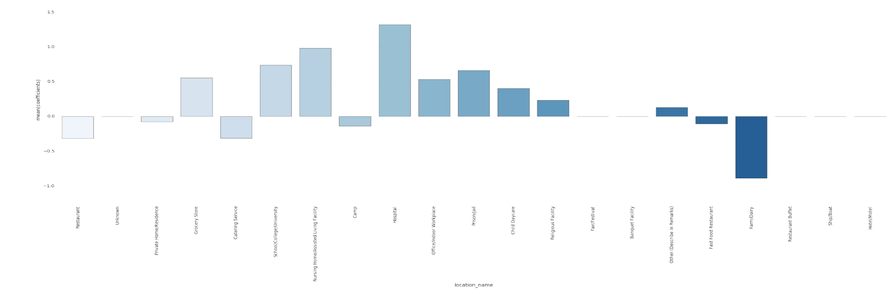
States



Food Source



Location

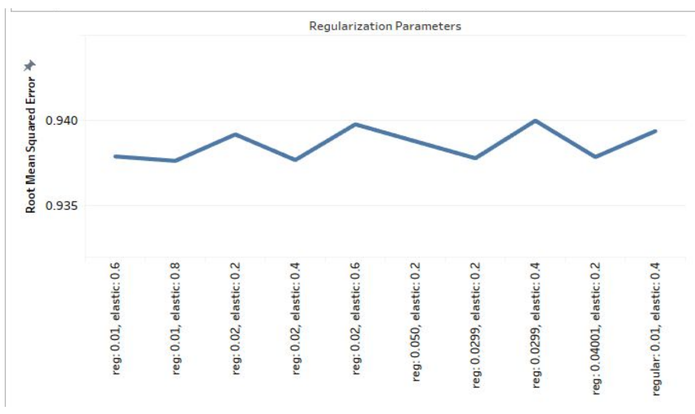


Model Description

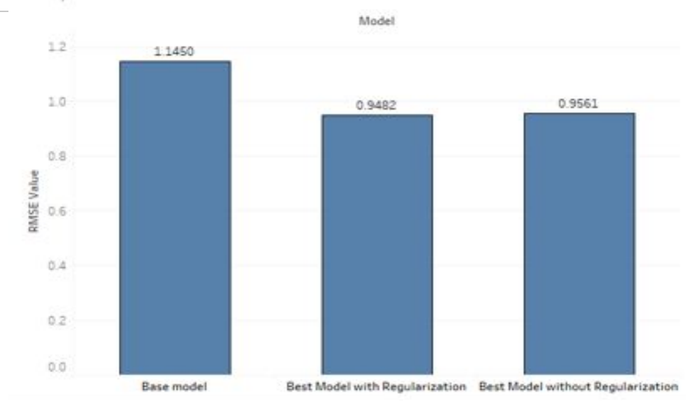
Model	Features	Techniques	Evaluation
Base Linear Model	Year	Spark pipelines	RMSE
Best Linear Model	Year, Month, State, location, Food	Spark pipelines, One Hot Encoding, String Indexer, Regularization, Cross Validation	RMSE
Logistic Regression	Year, Month, State, location, Food	Spark pipelines, Regularization, One Hot Encoding, Feature Scaling, Cross Validation	ROC curve, Precision, Accuracy, Recall value

Linear Regression model evaluation

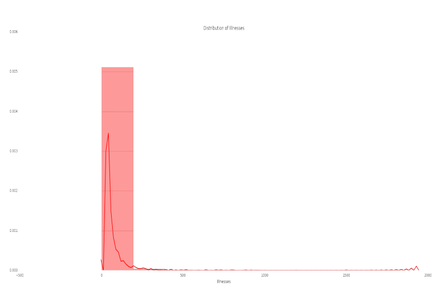
Model Evaluation

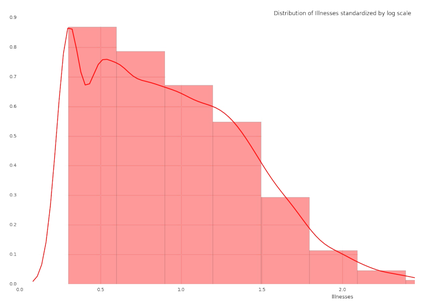


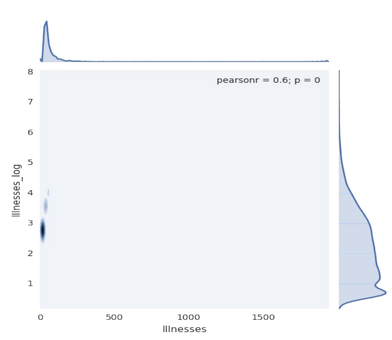
Model Comparison



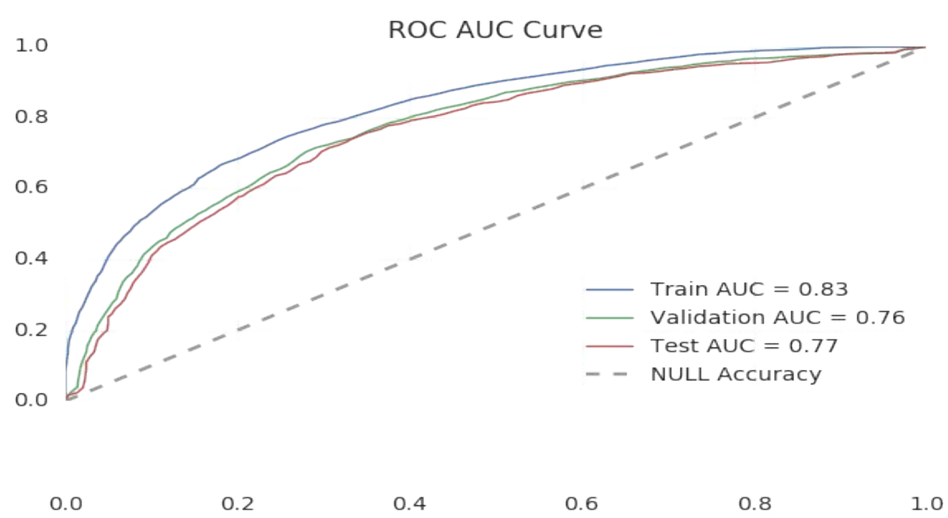
Prediction Value Normalization







Logistic Regression model evaluation



Conclusion

Use our platform to input year, month, state, Location and Food being consumed and in return our platform will forecast the amount of illnesses that can be produced based on inputs with an **RMSE of 0.948**. The platform also categorized illness as High or Low level with an **AUC of 0.77**. Lastly, the platform has a future scope of Time Series Forecasting to detect the illnesses trends in advance**. Stay Alert stay Healthy!**