

**MetLife Sinkhole Analysis**

Micheal Andrejco, Devon Bolen, Gina Catellier, John Leary, Caroline Williams

November 19, 2020

ISA 460

**Executive Summary:**

For more than a decade, the state of Florida has repeatedly experienced the highest number of sinkholes of any state in the nation. The reason for Florida’s sinkhole problem is due to the porous carbonate rocks that make up the state’s peninsula. Typically recognized as limestone, these rocks are what help groundwater move through the soil. On top of these rocks is a thin layer of soil, dirt, or clay. Overtime, the rocks begin to dissolve from acid created by oxygen in the water. Once rocks cannot support the weight of the soil above, the rocks collapse which forms a sinkhole. The problem has continued to persist over time which has made it difficult for insurance companies to cover high-risk areas.

To better assess where insurance companies such as MetLife should write policies or be more careful when writing policies in areas where sinkholes can occur, a machine learning model can be produced. The threat of the likelihood that a new sinkhole will emerge in a given area poses a threat to MetLife’s company when writing policies and can be addressed by creating models with features that incorporate all aspects of the environment where the sinkhole is occurring. By incorporating both natural and human factors into the model, a reliable and accurate prediction can be made in the occurrence of a sinkhole in a given location. The resulting model allows companies such as MetLife to update their policy standards and re-evaluate where or if they are willing to write policies in certain areas to help their company continue to be successful.

**Purpose of the Project:**

The goal of our analysis is to assess the likelihood of a sinkhole forming within a one-mile radius of a given location in Florida. We will be using a dataset with nearly 4.9 million sinkhole records collected by MetLife Insurance. We know that sinkholes are more likely to occur in certain areas based on the soil composition and new sinkhole emergence can be influenced by human factors as well. We aim to pull meaningful insights from the dataset about which features regarding the makeup of these areas can most reliably predict the occurrence of sinkholes in certain areas of Florida. After producing a model that can reliably predict an emerging sinkhole, recommendations will be made to MetLife on how they can move forward with writing policies in these areas.

**Approach:**

The first step in our analysis was to examine the raw data thoroughly. This would give us an idea of the size of the dataset along with the types of variables we would be using, and what patterns might be present in the data. The dataset consists of over 4.9 million rows containing information about Florida sinkholes. There are several categories that the features are broken down into. These include: Soil Characteristics, Census Data, Weather Impacts, and Historical Sinkhole Data. From this list, the categories with the most features happen to be data collected from the United States Department of Agriculture. This is primarily data which holds very specific information about soil characteristics at varying depths. There are also features within this category that provide insight into the rock type, water storage, and weather impacts. Other than topographical information, there are also location identifiers. These values are specific for the recorded addresses of the sinkhole reports. They include a specific geocode consisting of longitude and latitude points that are used as an identifier for the addresses in the data.

After exploring the data, some cleaning up did need to be done in order for the data to be usable in our models. Columns with more than 75% null values were removed from the dataset entirely. However if a categorical column had less than 75% null, that null value was replaced with the word ‘Unknown’ and if a numerical column contained nulls less than 75%, then the null was replaced with the mean of that specific column. In addition, outliers were also evaluated in the dataset and were defined as values that fall outside of three standard deviations. The last data cleaning that we did, involved changing the target variable into a binary target (will a sinkhole occur or not), instead of just having the target variable describe how many sinkholes occur within the one mile radius of a given location. The label of 0 was defined as no sinkholes occurring while the label of 1 implied that at least one sinkhole would occur.

Once the data was cleaned, feature selection and modeling was carried out. Feature selection can be broken up into different groups and requires us to test what groups, and in what amounts, do features work the best together to produce a precise model. In addition to this, the dataset is largely imbalanced so this must be taken into consideration when running the models and making predictions. The features and why we chose them will be discussed below in the model development section. In addition to working with different feature groups, different types of models are also important to test out during this project, such as logistic regression, random forest, and gradient boosted tree models, to name a few. It is unclear right away what would be the best model with the data, and thus multiple types have to be run with the features to see what initially provides the best results. The models we chose to run, and their results will also be discussed below including their specific results for precision and lift/gain values. It is important to have these results when determining the best model because prediction accuracy is not enough of an indicator to say if the model was reliable.

**Models Development & Tuning:**

In order to predict if a sinkhole will form within a one mile radius of a given location, feature selection in such a large dataset is key to providing a reliable model. As stated above, the data can be mainly broken down into groups such as soil, population, weather, and historical sinkhole data. With this information, three distinct feature lists were created to evaluate what feature combinations produced the best models. Feature list 1 consisted of aquifer data, some population data, some soil component data, and some precipitation data. Feature list 2 also had aquifer data but contained all population data, lots of soil data, and all precipitation data. Lastly, feature list 3 consisted of aquifer data, all population data, lots of soil component data, and very little of the precipitation data. A visual representation and more detailed explanation of these feature lists in the exhibits section of the report, specifically in Exhibit 1.

These features were selected because after doing background research, it was clear that these could possibly have the most effect on if a sinkhole were to occur. Aquifer and soil data were already known to be of importance when predicting sinkholes, but the addition of population/housing and precipitation data could tell us if a sinkhole is more likely to occur in certain areas. Population data is referring to total census block population for the prior year, the prior year population density change for a census block population, and the annual growth rate from 2010 to the prior year listed for each record. Housing data includes total prior year housing units for the block group, the housing unit density for the BG in the prior year from that listed, and the prior year housing unit density change for the block group.

The aquifer and soil data that was used in the various feature lists always included aquifer name and some varied amount of data on available water storage (AWS), thickness of soil components (TKO), soil organic carbon stock estimate (SOCO), and data from the national commodity crop productivity index (NCCPI). These features were in each of the lists, but as stated above varied in how much and how detailed they were in inclusion in certain lists. Thus, if one feature list says ‘some soil component data’ vs another list that says ‘lots of soil component data’, it can be inferred that the feature list with lots of soil component data is more detailed than the other. In addition to these features being included in each feature set, the feature that described if there already was a prior existing sinkhole was also included in each set because it is well known that if a sinkhole already exists, another sinkhole is more likely to appear in that area.

**Modeling Approach:**

As discussed previously, the dataset was very imbalanced. There were far more records of zero sinkholes within a one mile versus than any at all. It was due to this imbalance that we decided to oversample the training data. In theory this would allow us to achieve a more balanced sample to train our model on. Without oversampling, the training set would not have enough records of a sinkhole happening to make a prediction for what is causing them. Before oversampling, our target variable ratio of sinkholes to no-sinkholes was 0.03%. Because of this, we duplicated 40% of the minority class (non-sinkhole class) relative to the data as a whole. This left us with an ending ratio of 0.4%. After this, we then recreated our training data set to be used in the modeling and prediction phase of our research.

To produce the best model for this project, multiple types of models were tried to see what type produces the best result. The models that were initially implemented included Random Forest, Logistic Regression, and Gradient Boosted Tree. We decided to use these models first because they would be the easiest to run with the data we had. For example, logistic regression was chosen since the target of our analysis was whether or not a sinkhole will occur. Logistic regression performs well with these types of features since the output is defined as either a zero or one. However, we felt that using a model that incorporates multiple decision trees will be more accurate as the models are combined either in parallel or boosting to prevent overfitting. Random Forest was another model we used because it trains each tree independently using a random sample of the data, making the model more robust and less likely to be overfitted. However, we felt that gradient boosting would be the best model to use on our data due to its ability to classify trees and adjust the weights, allowing the model to improve based on prior predictions.

**Model Performance:**

After running the three different types of models, our logistic regression and random forest models produced poor precision and lift results and it was determined that Gradient Boosted Tree models had the best numerical results based on precision and lift. Each feature lists results for gradient boosted tree models are listed below:

* Feature List #1:
  + Precision: 89.5
  + Lift: 6.46
* Feature List #2:
  + Precision: 86.18
  + Lift: 6.45
* Feature List #3:
  + Precision: 98.78
  + Lift: 5.07

Feature lists one and two show similar results with precision and lift, thus giving the idea that the differences between the two feature groups did not have a big enough impact on the results. However, feature list 3 stood out in precision from the rest but did have a lower lift value than the other two sets. The actual lift and gain plots for feature list 3 can be seen in Exhibits 2 and 3. After running cross validation for feature list #3, the lift was greatly increased to 6.48 for docile group 1.

**Discussion of Results:**

After a preliminary analysis it was clear that the Gradient Boosted Tree models were producing the best results. This model was run with all three feature sets the team had chosen. Initially feature sets with just soil and aquifer data were run, but afterwards data regarding weather and population was added to the list of features used in both models. This addition of weather and population data significantly increased the prediction accuracy of our models and proves that the prediction of a sinkhole cannot just be based on soil data. While running our models, it was evident that this was true based on the feature weight importance calculations that were carried out (Exhibit 4). The feature importance weight calculations were able to tell us that the top weighted features for feature list #3 included annual precipitation, prior existing sinkhole count, population density percent change, and housing density percent change. These can be seen in the exhibitions page.

This also reaffirms some of the cases plotted before the modeling process had even begun. Many of the high-risk areas for sinkholes are around coastlines. Areas such as Tampa and St. Petersburg are prone to rock erosion due to the harsh weather effects brought on from the gulf coast. Even inland cities such as Ocala repeatedly witnessed a steep amount of sinkholes over the last four years. This can be seen by taking the original data points of reported sinkhole occurrences and plotting them against their actual locations in Florida.

**Recommendations:**

Our findings could greatly impact the policy holders in certain regions of Florida. Based on our model performance, it would be beneficial for MetLife to not write any policies for docile groups one and two, based on the lift chart results that indicate there is a higher chance a sinkhole will occur (Exhibit 5). We feel that all of the other docile groups listed can have policies written for them. For current policy holders this could influence how they move forward with MetLife, depending on if they are in an area with a high chance of a sinkhole emerging.

While we have made these recommendations based on our model and findings, it must be noted that the dataset could be flawed in ways that were out of our control to handle. For example, *The Florida Times Union* says that state insurance officials believe many claims could be false. For a policyholder to claim a loss, there must be structural damage to a home, but many claims are paid without that proof. This would mean the actual number of sinkholes is less than what gets reported. If true, this would have skewed our data and overall recommendations, as we might be reporting that a sinkhole be more likely to occur somewhere where there has only been false claims, and not any actual sinkholes emerging.

In addition to this, the dataset could be biased toward certain areas or locations in Florida that have the ability and funds to report the occurrence of a sinkhole. If this is true, there could be large parts of Florida that do in fact have occurrences of sinkholes but are missing from the dataset. This could affect the overall performance of our model because we would be missing the data linked to those sinkholes and some areas that we reported as low risk may in fact not be.

As a result of completing this project, MetLife should consider re-evaluating what areas of Florida they write policies for, given the fact that some areas are more susceptible to emerging sinkholes. These sinkholes can emerge from a combination of natural factors such as aquifer, soil composition, and any prior existing sinkholes, but can also be influenced by human and weather factors. As precipitation patterns change and affect the soil, sinkholes can become more likely and the population of people that are living in certain areas can result in an increased chance of sinkholes occurring. While this data can be slightly skewed by reporting flaws as described above, it is clear that certain areas in Florida have a higher risk and therefore should be under different policy conditions for MetLife to ensure that they are making the best decisions for the company.

**Challenges & Lessons Learned**

While completing this project there were some challenges that we faced as a group. The first roadblock we faced was when trying to find a way to break down the analysis of the project for each member in the group. Since the dataset is so big and contains so many features, it was sometimes difficult to explain each step of our work to the other team members. It was also very important that each team member looked over each other’s code to make sure it was correct. This was a critical step because some outputs may look correct but small differences could result in mistakes.

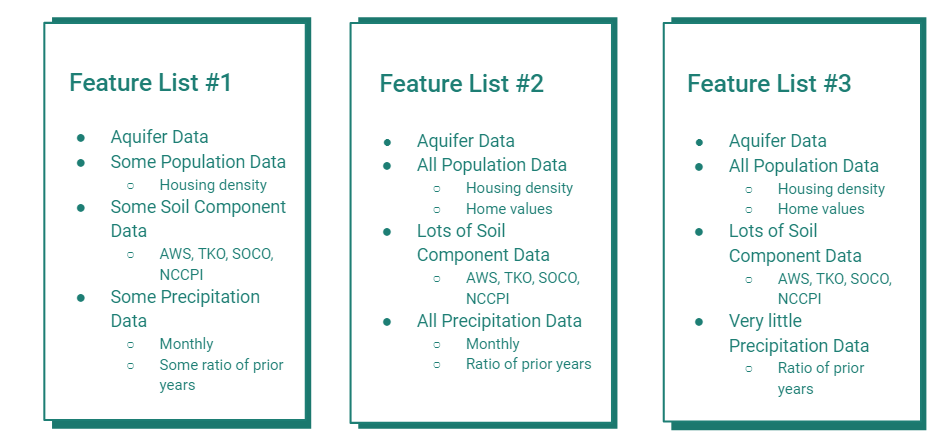
Furthermore, we had trouble finding meeting times as a group that worked for everyone. Since the semester was so condensed and our group members are involved in athletics, work-study, organizations, other group projects, and club sports it was difficult to find time that worked for everyone. Also, since we had so many time constraints a majority of our meetings were over zoom and some proved to be inefficient. Reflecting on this we realized we should have sent agendas for our meetings as well as assign a member to head the meeting and keep the group on task. Another challenge we faced was time management. For example, the condescended semester created an obstacle in which our group thought that we had more time then we really did. This created unneeded stress when deadlines were fast approaching and caused our group to have to work under pressure. Finally, the biggest challenge we faced was finding a way to work on the notebook for the models together. The shared notebooks seemed to be clunky and not user friendly when multiple group members were using the same notebook. This made it difficult to put together one notebook with everything that we needed for the project.

After completing this project, we now have a better understanding of the whole process of starting with a dataset and transforming it to run and refine a machine learning model using Pyspark. To create the best model for this project a lot of revisions to our feature lists and our models had to be made. Since the best model depended on our data and feature lists a lot of different types of models had to be run which allowed us to learn how to code each for our dataset. In addition, learning the role of lift and gain charts allowed us to better understand our results and put the results of our model into words. The culmination of these objectives are what allowed us to make accurate predictions for a large company like MetLife. Also, as a group we had almost no experience prior with machine learning so this semester was not only about learning how to code the various models with pyspark, but understanding what the model is trying to accomplish.

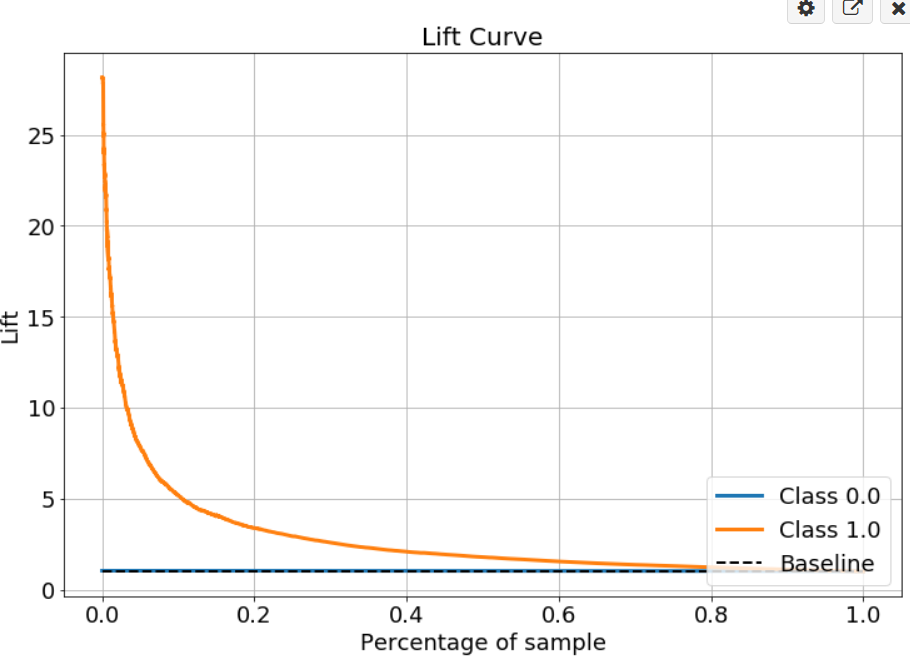
Overall, as a group, we have become more competent in not just building, but understanding machine learning techniques and models. Finally, the most important lesson learned was how to work together efficiently as a group and utilize each individual member's strength while minimizing their weaknesses. It started to become apparent throughout the semester which group member excelled at doing what and became easy to assign each member to a different task. In conclusion, this helped us to submit a report and model that was effective in predicting a sinkhole.

**Exhibits/Visualizations**

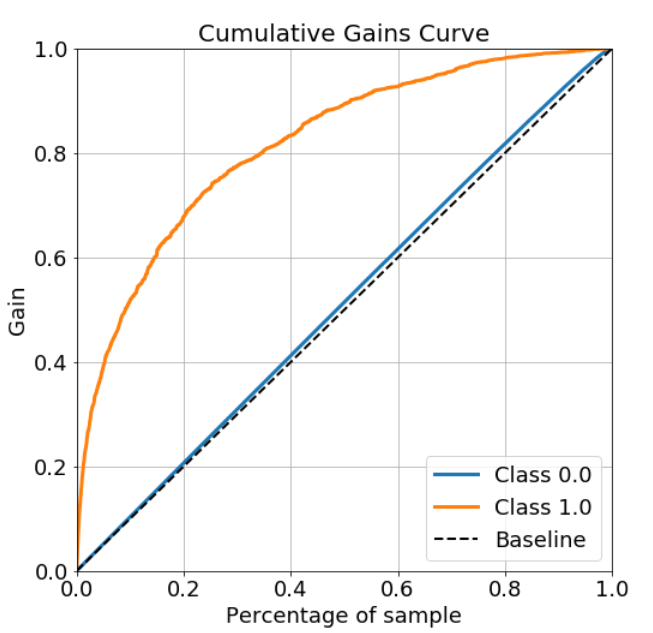
**Exhibit 1**



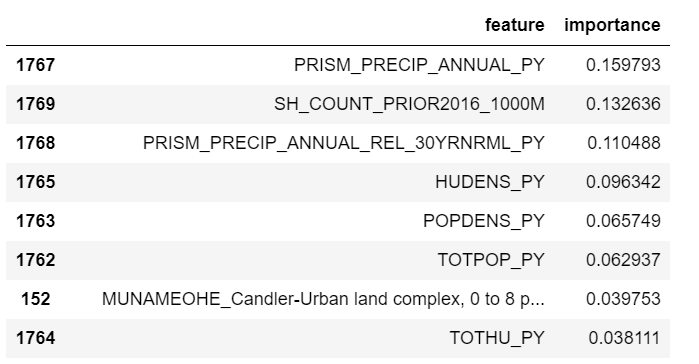
**Exhibit 2**



**Exhibit 3**



**Exhibit 4**



**Exhibit 5**

