|  |  |
| --- | --- |
| Regression ANalysis on monetary Loss from house fires | By: Tyler Bier |

**Research Question**

In data (2018-2022) published by the Insurance Information Institute evaluating insurance company losses, fire and lightning (coded together) make up the highest cost per claim among property damage claims at $83,991. For comparison, the second highest property damage claim cost is water damage and freezing, at $13,954. Within the overall insurance industry, fire and lightning were responsible for 21.9% of all homeowner insurance policy claims in the year 2022 (Insurance Information Institute, 2024). Summary statistics published by the U.S. Fire Administration estimated that there were 374,300 residential building fires nationally for the year 2022. In turn, causing “$10,821,300,000 in dollar loss” (U.S. Fire Administration, 2022).

With insurance companies facing potentially significant costs due to residential fire claims, the industry adopted the Public Protection Classification (PPC) program to determine appropriate premium rates considering fire risks. “The Public Protection Classification (PPC) program is a community fire protection scoring system based on a Fire Suppression Rating Schedule (FSRS) used by the Insurance Services Office, also known as ISO Mitigation... Most home and business insurers use PPC scores to determine insurance premiums. Businesses and individuals in a community with a good PPC score typically pay less for fire insurance” (Public Protection Classification, 2015). The PPC program gives communities a rating ranging from 1-10 as a measure of local fire protection. This rating is determined by points allocated based on the above-mentioned Fire Suppression Rating Schedule (FSRS). The number of points available for allocation is generally 1-100, although this can slightly vary by state. The FSRS distributes points based on community fire department capabilities and metrics. Some of these determining factors include but are not limited to: Emergency dispatch center capabilities, fire department training, equipment, number of fire engines, mutual aid/ assistance agreements with neighboring fire departments, number of staff, and available water supply.

While the insurance industry has managed to successfully implement a graded system for overall community risk, it does not have the ability to predict or forecast actual potential monetary losses once a fire was to occur. Additionally, there are many other potential risk factors and variables that it does not consider that could potentially play a large role in financial costs from fires. To further explore this, we will analyze internal fire department incident data to identify factors affecting monetary costs and create a predictive model estimating financial costs of residential house fires. We will achieve this via various models of regression and other statistical testing.

**Our research question for this analysis is as follows:** Can a predictive regression model be constructed to estimate monetary loss from house fires that shows predictive improvement between multiple linear regression, stepwise regression, or random forest regression?

**Null Hypothesis (H₀):** A predictive regression model cannot be built on the selected dataset.

**Alternative Hypothesis (H a):** A predictive regression model can be built on the selected dataset.

**Data Collection**

For this analysis, data was obtained from the National Fire Institute Reporting System, commonly referred to as NFIRS. NFIRS is a dataset containing each individual emergency incident and details about said incident from reporting fire departments on an annual basis. “Nationally, NFIRS is used by various private industries, including national associations for home appliance product manufacturers, the hotel and motel industry, insurance companies, and attorneys” (U.S. Fire Administration, 2015). This is in addition to federal agencies. NFIRS data is a collection of 19 relational data tables by year. It is estimated to capture ~70% of fire data in the USA. Each table “collects a common set of information that describes the nature of the call, the actions firefighters took in response to the call, and the end results, including firefighter and civilian casualties and a property loss estimate” (U.S. Fire Administration, 2024). NFIRS is managed by the U.S. Fire Administration, who in turn are under FEMA. Specific dataset used is ‘2022 public release’ and is freely accessible for the public domain. All individual identifying information has been removed from public release version.

An advantage to using the NFIRS dataset on fire related analysis is that it provides unique insights into actions taken by firefighters and fire investigators during an incident. Examples of such insights include response times, actions taken by firefighters, release of hazardous materials, number of fire apparatus used in an incident, firefighter injuries, civilian injuries, and whether mutual aid assistance was required of neighboring departments. Data points such as these cannot be found in datasets that are outside fire departments.

A disadvantage of the NFIRS dataset is that the data is input directly by firefighters and fire investigators working from the field. When considering such work conditions from the field, this will likely result in a certain degree of human error within the dataset. Such as incorrect estimates, and some values entered incorrectly or not recorded at all. Due to this, the nature of this dataset should work well with determining overall fire related estimates at a national or state level. At the same time, it should not be used for exact quantitative approximations of certain data points at the national/ state level. As there is that degree of human error and only ~70% of national fire data is collected here. This element of human error is something that will be heavily considered when cleaning the dataset or considering use of certain variables.

The largest challenge with using the NFIRS dataset is simply the scale of it. The dataset contains a set of 19 relational .txt data tables. One of the tables used for this analysis, ‘causes’, is one of the smaller tables and contains 320,088 rows with 9 columns. Larger tables such as ‘basicincident,’ have over 30 million rows, with 42 columns. To remedy this challenge, the focus of our research question and careful selection of chosen variables significantly reduced this scope to make this analysis more manageable. By choosing to ask if estimating monetary losses from residential fires with regression is possible, the number of tables required for such analysis is drastically reduced. Three tables from the overall dataset will be used for this analysis. Tables named ‘basicincident’,‘fireincident’ and ‘causes.’ After that reduction however, there is still a vast number of variables to consider. Between these four tables, there is a combined 132 columns. With many of these columns being variables unique to their respective table. Each table has a range of 4-7 variables that are shared amongst other tables. Such as one unique primary key shared amongst all tables, and other variables representing key fields shared amongst selected tables. When considering which variables to include or exclude, there is several practical resources included the download of the NFIRS dataset. An important one being the National Fire Incident Reporting System complete reference guide. “It provides both instructions for reporting data to NFIRS Version 5.0 and an understanding of the data elements collected by the system. It also serves as a reference for the coding of the data” (U.S. Fire Administration, 2015). This reference guide explains the measure of each variable and explains numerical codes representing different categories for each variable. As most categorical variables have different unique numerical codes representing respective categories or levels. An example being 419 and 421 representing distinct types of residential structures. By cross referencing this reference guide with our selected tables, it allowed reduction by determining whether certain variables were relevant to our research question or whether the quality of certain variables were appropriate for this analysis. Some variables within the dataset were not required to have data input, resulting in a large amount of missing or incomplete data in some variables that was not appropriate for this analysis. To avoid this, most of the variables chosen were those that were required to have data input. To reiterate, a focused research question and careful selection of variables using supplemental resources allowed us to filter our data and overcome the massive scale of the overall NFIRS dataset.

A selection of chosen variables for this analysis can be found in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Type of Data** | **Role of Variable** | **Role** |
| estimated\_loss | Continuous | Dependent; Created; Sum of Prop\_loss & Cont\_loss | Monetary loss ($) |
| response\_time | Continuous, Time/ Date | Independent; Created; Result of Alarm - Arrival | Time to arrive at scene in min |
| STATE | Categorical | Independent | State |
| FDID | Categorical | Independent | Code representing various FD |
| INC\_DATE | Continuous, Time/date | Independent | Date of incident |
| AID | Categorical | Independent | Whether aid was given/ received from other FD |
| ALARM | Continuous, Time/date | Used to calculate response\_time | When call was received/ dispatched |
| ARRIVAL | Continuous, Time/date | Used to calculate response\_time | When responding unit arrived on scene. |
| SUP\_APP | Continuous | Independent | Number of fire apparatus on incident |
| PROP\_LOSS | Continuous | Used to calculate estimated\_loss | Estimated loss amount of property in $ |
| CONT\_LOSS | Continuous | Used to calculate estimated\_loss | Estimated loss amount of contents in $ |
| PROP\_USE | Categorical | Independent | Type of residential property |
| AREA\_ORIG | Categorical | Independent | Area where fire originated |
| DETECTOR | Categorical | Independent | Fire detector presence |
| AES\_PRES | Categorical | Independent | Automatic extinguishers presence |
| CAUSE\_CODE | Categorical | Independent | Cause of fire |
| STRUC\_TYPE | Categorical | Independent | Type of structure |
| STRUC\_STAT | Categorical | Independent | Occupancy status of structure |
| BLDG\_ABOVE | Continuous | Independent | Number of stories |
| CONF\_ORIG | Binary | Independent | Fire contained to origin point |
| FIRST\_IGN | Categorical | Independent | First item ignited |
| HEAT\_SOURC | Categorical | Independent | Heat source that led to ignition |
| FLAME\_SPRD | Binary | Independent | Did fire spread? |
| FIRE\_ORIG | Continuous | Independent | Floor level/ story that fire originated |
| INC\_TYPE | Categorical | Independent | Type of fire incident |
| ACT\_TAK1 | Categorical | Independent | Action taken by personnel |
| HAZ\_REL | Categorical | Independent | Were hazardous materials released? |
| TOT\_SQ\_FT | Continuous | Independent | Square feet of building |
| FF\_DEATH | Continuous | Independent | Fire service deaths |
| OTH\_DEATH | Continuous | Independent | Civilian deaths |
| FF\_INJ | Continuous | Independent | Fire service injuries |
| OTH\_INJ | Continuous | Independent | Civilian injuries |

Link to NFIRS dataset will be found in sources.

**Data Extraction and Preparation**

For data extraction, preparation, and all other stages of this analysis we will be using R within R Studio as our environment. An advantage of R versus another language like python is “R programming is better suited for statistical learning, with unmatched libraries for data exploration and experimentation” (IBM Cloud Team, 2024). Examples of such libraries are packages like ggplot2 for distinct visualizations, and caret for quick manipulation and statistics on machine learning models. Disadvantages of R when compared to python, is that the performance of R is primarily limited to CPU cores. Resulting in slower speed when training models, among other functions. Taking advantage of GPU performance on R requires several packages, external downloads, and other factors that need to be manipulated. Whereas python has several packages that allow for much quicker and streamlined GPU usability. Python is also much more streamlined when it comes to integrating with cloud computing resources, such as Azure.

Next, we will begin the data-extraction and -preparation process step by step. This will start with loading our four chosen tables into the R studio environment, and merging these tables based on their common primary key “INCIDENT\_KEY.” Undesired variables will then be removed. Following will be standard data cleaning processes. Any duplicate rows will be removed. Missing/ Null values will be imputed as deemed appropriate by underlying data distribution. To get dataset to a more manageable size, data will be filtered down among the variables PROP\_USE, INC\_TYPE, and AID, as recommended by the NFIRS 5.0 Fire Data Analysis Guidelines and Issues document for analyzing residential structure fires (FEMA, 2011). Outliers will be examined with treatment being removal, imputation, or to remain based on specific variable and data distribution. Variables PROP\_LOSS and CONT\_LOSS will be summed into new variable, estimated\_loss. Estimated\_loss will be our dependent variable that will represent financial loss from residential fires. Variables ALARM and ARRIVAL will be differenced to create new variable, response\_time. This variable will represent travel time from fire department dispatching personnel to arrival at scene of incident. Categorical variables will be examined, with those having excessive category amounts having those categories reduced to smaller, respective groupings. To make later modeling less computationally taxing.

Note: All code showing the exact steps needed for data extraction and preparation are attached. Code can also be seen in its entirety in submitted R notebook.

Beginning with loading and merging chosen three tables into R studio based on their primary key. Two initial packages loaded are tidyverse for its data manipulation packages and ggplot2, and then the data.table package. The data.table package is essential for importing the larger data tables into R studio, as base R import packages have difficulties due to size of NFIRS tables.

A computer screen shot of a computer code

Description automatically generated

A screen shot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated

After combining tables into fireincidenttable, will now select for our earlier specified variables and the primary key.



Checking for duplicated rows, none are found.

A screen shot of a computer

Description automatically generated

Checking for NA values. See below.

A screenshot of a computer

Description automatically generated

Checking for missing data. See below.

A screenshot of a computer

Description automatically generated

There is a large amount of NA and missing values across multiple columns. However, a large portion of these NA and missing values are likely because they are tied to incidents that do not involve residential house fires. So, before cleaning/ imputing values for either missing or NA, we should filter our available data as referenced above in order to ensure our current dataset is limited to residential structure fires.

Filtering AID, INC\_TYPE, and PROP\_USE based on recommendations from previously mentioned NFIRS 5.0 Data Analysis Guidelines and Issues.

Filtering AID rows 3-4.A screenshot of a computer

Description automatically generated

Filtering INC\_TYPE via various residential fire code categories and confirming replacement.

A black screen with text

Description automatically generated with medium confidence

Filtering PROP\_USE on 419 only. 419 is coded for: 419 1- or 2-family dwelling, detached, manufactured home, mobile home not in transit, duplex. Which is our desired target for residential house fires. We will then remove the column to reduce redundancy.

A screenshot of a computer program

Description automatically generated

Also going to be filtering on STRUC\_TYPE on category 1, which captures enclosed buildings. All other categories are not relevant for analysis, will then remove column.

A screenshot of a computer program

Description automatically generated

Rechecking NA values after filtering for only residential house fires.

A screenshot of a computer

Description automatically generated

Rechecking missing values.

A screenshot of a computer

Description automatically generated

Addressing PROP\_LOSS NA values and outliers. Summary statistics below, histogram and boxplot show data distribution is heavily skewed towards the right. Imputed values for NA will be the median at 4000. Outliers will be identified via scaling the variable and imputed with the median if an outlier has a scaled value of 3 standard deviations away from the mean.

A graph with numbers and lines

Description automatically generated A white rectangular object with black lines

Description automatically generated

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0 0 4000 42827 30000 100000000 3330

A screenshot of a computer

Description automatically generated

Addressing CONT\_LOSS column, values and outliers. Data showing another right sided skew, with a fair number of outliers and missing values. Imputed values for NA will be the median at 400. Outliers will be identified via scaling the variable and imputing with the median if an outlier has a scaled value of 3 standard deviations away from the mean.

A graph with numbers and lines

Description automatically generatedA white rectangular object with black lines

Description automatically generated

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0 0 400 12730 10000 2000000 3825

A screen shot of a computer

Description automatically generated

Addressing missing values within variable SUP\_APP. Right sided skew, NA values and outliers present. Imputing median for NA values and outliers, based on z-scores of 3, -3.

A graph with numbers and a line

Description automatically generated A white sheet of paper with black lines and dots

Description automatically generated

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0.000 2.000 4.000 5.568 6.000 7200.000 2903

A screen shot of a computer code

Description automatically generated

Addressing outliers in BLDG\_ABOVE variable. Data is skewed right with a median value of 1. Imputing outliers with median value.

A graph with numbers and lines

Description automatically generated A white rectangular object with black lines

Description automatically generated

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000 1.000 1.000 1.441 2.000 324.000

A screen shot of a computer

Description automatically generated

While reviewing the variable CONF\_ORIG for NA values, it is determined that both CONF\_ORIG and FLAME\_SPRD appear to be both binary variables capturing whether fire left its origin point. To reduce redundancy, will be removing variable CONF\_ORIG as data is captured in another variable.



Addressing variable FIRE\_ORIG for outliers. Will be imputing outliers with median of 1 due to right sided skew of data. Outliers determined by z-scores based on distance of 3 standard deviations.

A graph with numbers and a bar

Description automatically generated A white paper with black lines and dots

Description automatically generated

Min. 1st Qu. Median Mean 3rd Qu. Max.

-4.000 1.000 1.000 1.228 1.000 950.000

A screen shot of a computer

Description automatically generated

Addressing NA values and outliers within TOT\_SQ\_FT column. Both NA values and outliers as determined by z-scores will be imputed with median value due to right skew of data.

A graph with numbers and lines

Description automatically generated A white rectangular object with a black line

Description automatically generated

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0 1000 1400 2230 1989 28600000 17422

A screen shot of a computer code

Description automatically generated

When assessing NA values in FF\_DEATH, we can see the number of actual deaths is very low compared to the overall data. With only 3 deaths among entire dataset, this variable will be removed due to low observations.

A screenshot of a computer

Description automatically generated

A grey background with white text

Description automatically generated

Next column is OTH\_DEATH, will be imputing missing values and outliers based on z-scores with our median of zero.

A graph with numbers and a bar graph

Description automatically generated A diagram of a diagram

Description automatically generated

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0.00 0.00 0.00 0.36 1.00 6.00 78866

A screenshot of a computer program

Description automatically generated

Once correcting for NA values, all civilian deaths were considered outliers. Will be removing this variable.



Addressing NA values in FF\_INJ. Which number at two.

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0.00000 0.00000 0.00000 0.03334 0.00000 24.00000 2

With a current data frame of 81,649 rows, this is another variable with too little observations. Removing it from the data frame.



Next column with NA values is OTH\_INJ. Missing values will be imputed with median of 1. Outliers will be assessed and imputed with 1 if they are outside standard z-scores.

A graph with numbers and a bar

Description automatically generated A white rectangular object with black lines

Description automatically generated

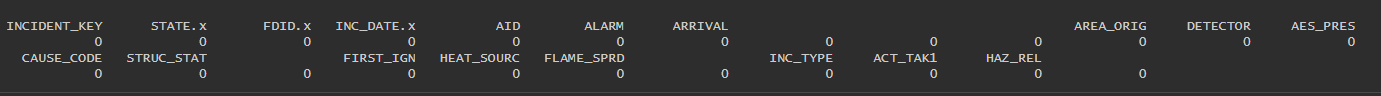
Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0.00 0.00 1.00 0.91 1.00 9.00 78866

Due to the large number of missing values overall, this variable will be removed.



All NA values have been addressed.



Addressing missing values in HAZ\_REL, a categorical variable. Table() function below. Will be imputing missing values as the mode, N.

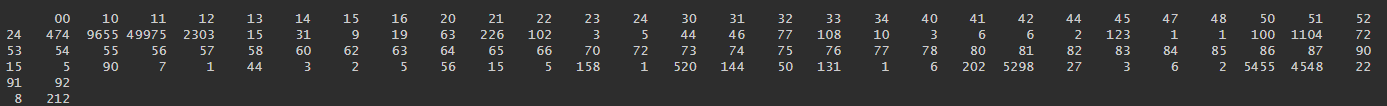
0 1 2 3 4 5 6 7 8 N

48438 25 39 76 54 4 13 17 8 4 32971

A screenshot of a computer

Description automatically generated

Addressing missing values in ACT\_TAK1, a categorical variable. 24 missing values are showing. Will impute them with the mode, 11.





Now confirming NA values and missing values are removed throughout dataset.

A black screen with white text

Description automatically generated

Moving forward with data transformation, ARRIVAL and ALARM variables will be converted into a date/ time format, then ARRIVAL will be subtracted from ALARM in order to create a new variable, RESPONSE\_TIME. Which will indicate travel time from fire station to scene of incident in minutes.

A screen shot of a computer program

Description automatically generated

Next will evaluate and impute potential outliers in this column.

A graph with numbers and a line

Description automatically generated A white rectangular object with black lines

Description automatically generated

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000 4.000 6.000 7.631 9.000 6360.000

Outliers will be determined via z-score metrics of 3 standard deviations away from mean. Right data skew indicates imputation of the median.

A screenshot of a computer

Description automatically generated

For RESPONSE\_TIME, since you cannot have a realistic response time of zero minutes, values of zero will also be imputed with the median.

A screen shot of a computer

Description automatically generated

Now creating the column/ variable ESTIMATED\_LOSS, which will be the dependent variable for later models. It will be created as a sum between PROP\_LOSS and CONT\_LOSS and be indicative of total losses from residential house fires.

A screenshot of a computer

Description automatically generated

Due to a strong right sided skew in the newly created ESTIMATED\_LOSS. Outliers will be evaluated based on previous z-score metrics and replaced with median of 4400.

A screen shot of a computer

Description automatically generated

Within the INC\_DATE variable, which represents month/day/year, we will remove elements so only months are represented within our analysis.



Removing redundant variables used in earlier variable creation.



The categorical variable FDID has 9,208 categories within. To remedy this for later modeling, we will place these categories into one of three categories based on number of recorded incidents they have via data splits at .33 and .66. Labels will be called “Small Dept”, “Medium Dept”, “Large Dept.” These categories should correlate with the size of the department based on number of responses made.

A screenshot of a computer program

Description automatically generated

Setting variables as either factors or numeric.

A screen shot of a computer program

Description automatically generated

Removing additional columns.



With some of the other categorical variables in this analysis having a high number of overall categories, we will be reducing those categories into similar groupings. Starting with ACT\_TAK1. Specifically, similar groups such as those involving Fire control, Informational investigation/ large incident command, and Mutual Aid.

A group of colorful text boxes

Description automatically generated with medium confidence

Doing the same with FIRST\_IGN, which records first item ignited in a fire. These will be grouped into Structural Component, Furniture, Bedding/ Clothes, Recreational Material, Storage Supplies, Liquids/ Pipes, Organic Mats, general\_mats, and Undetermined.

A screenshot of a computer screen

Description automatically generated

Because the context of this analysis is to predict estimated loss from house fires in regard to insurance company homeowners’ insurance claims data, we will be filtering out all rows where estimated loss is at zero. As there would not normally be a claim of zero dollars in losses.



**Analysis**

For this analysis, several regression techniques will be used on our current available data to determine prediction capabilities. “Regression is a statistical approach used to analyze the relationship between a dependent variable (target variable) and one or more independent variables (predictor variables)” (Shukla, 2017). Specific regression techniques used include multiple linear regression, stepwise regression, and random forest regression. These techniques performance will be measured and compared to see if any offer increased predictive accuracy.

Multiple linear regression is a relatively quick and simple regression technique to implement. Where the model will attempt to find coefficients and an intercept line that best fits our dataset. However, multiple linear regression makes multiple assumptions does make multiple assumptions of the data are met. If these assumptions are not met, it could result in a model that is not might be reliable. These assumptions include that there is a linear relationship between our dependent variable and independent variables, that there is not a high degree of correlation between independent variables used in the model, that variables have values that are independent of each other, presence of homoscedasticity( residuals of the model have similar variance throughout the results of the model), and that the data distribution of our dependent variable is normally distributed.

With stepwise regression, predictor variables are removed or added to a regression model until a model of best fit is found based on various criteria. Our stepwise regression model will specifically be measuring based on AIC score, which is a measure based on better overall model fit with lower predictors needed for said improved fit. An advantage to stepwise regression is that the process of adding/ removing predictor variables is automatically handled via several packages within R. A disadvantage of stepwise regression is that because variables are chosen by the model based on only one specific criteria (AIC in our use case), stepwise regression can result in “overfitting the data, biased estimates, and inflated Type I error” (Ravelo, 2022).

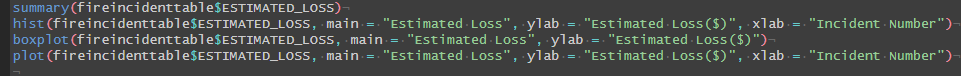
“Random forest is one of the most popular algorithms for regression problems (i.e. predicting continuous outcomes) because of its simplicity and high accuracy” (Keboola, 2020). A disadvantage of training random forest models is that they “can lead to greater memory utilization and training times, especially on systems with limited resources” (GeeksForGeeks, 2024). This is due to the model creating/ calculating a large number of separate decision trees, which in turn must be stored on memory. The mean is then taken from all trees created and then used to make predictions.

To begin an analysis, we will first observe our overall dataset that will be used in predictive regression modeling after preparation and cleaning.

A screenshot of a computer

Description automatically generated

Next, will be evaluating data distribution of our dependent variable: ESTIMATED\_LOSS.



Min. 1st Qu. Median Mean 3rd Qu. Max.

1 4400 17500 64522 72000 1600000

A graph with a bar graph

Description automatically generated A graph of a graph showing a long thin line

Description automatically generated with medium confidence

The value range of estimated loss is 1:1600000. There is a strong right sided skew to the data distribution. As noted earlier, multiple linear regression assumes that the underlying data will be normally distributed. Otherwise, predictive models could be unreliable. To remedy this, a log transformation will be made on the data prior to modeling in order to achieve a more normalized data distribution.

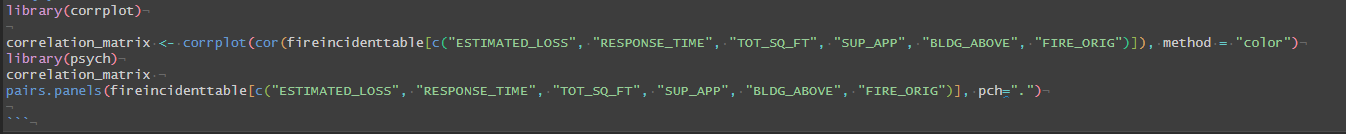
Next will be evaluating correlation and verifying linear relationships exist between our dependent variable, ESTIMATED\_LOSS, and our various numerical, continuous predictor variables. Correlation plots, the associated data distribution plots, and correlation scores shown below.

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated



Using the above visualizations, none of the above variables have a recorded correlation score of 0.6<, or -0.6>, indicating there should not be a degree of correlation high enough to remove said variables from modeling. Linear relationships of varying strength can be identified with ESTIMATED\_LOSS and predictor variables.

As data distribution of ESTIMATED\_LOSS is not normalized, parametric tests like ANOVA will not be used. Will use non-parametric tests on categorical variables. Below Kruskal-Wallis test will be used on categorical variables. Kruskal-Wallis test aims to identify differences in medians between various groups that are statistically significant. Any variables that do not have a statistically significant p-value result of test (>0.05) will be removed from any potential modeling. Results are below:

A screen shot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

All variables have a p value that is statistically significant, indicating there is a significant difference of median between the various grouping within categorical variables when compared with the dependent variable. All above variables will be retained.

Following are boxplot visualizations of the various categorical variables, and their underlying data distributions when measured against ESTIMATED\_LOSS.

A screen shot of a computer program

Description automatically generated

AREA\_ORIG: The area in which a fire originated.

A graph of different sizes and colors

Description automatically generated with medium confidence

CAUSE\_CODE: Cause of fire

A graph of a graph

Description automatically generated

FIRST\_IGN: First item ignited.

A graph of a graph

Description automatically generated with medium confidence

FLAME\_SPRD: Did fire spread?

A graph of a graph showing a number of different types of data

Description automatically generated with medium confidence

HAZ\_REL: Hazardous materials involvement.

A graph with a number of boxes

Description automatically generated with medium confidence

FDsize: Size of fire department.

A graph of a number of objects

Description automatically generated with medium confidence

INC\_DATE.x: Month that fire occurred.

A graph with black dots

Description automatically generated with medium confidence

INC\_TYPE: Type of incident responded to.

A graph with numbers and lines

Description automatically generated

Detector: Are fire detectors present?

A graph of a number of objects

Description automatically generated with medium confidence

AES\_PRES: Are automated fire extinguishers present?

A graph with a number of lines

Description automatically generated with medium confidence

STRUC\_STAT: Current use of structure.

A graph with lines and dots

Description automatically generated

AID: Was mutual aid utilized?

A graph of a graph with a number of black and white bars

Description automatically generated with medium confidence

HEAT\_SOURC: Source of heat.

A graph of different sizes and colors

Description automatically generated with medium confidence

ACT\_TAK1: Actions taken by responders.

A graph with numbers and a line graph

Description automatically generated

STATE.x: State that fire occurred in.

A graph of numbers and letters

Description automatically generated with medium confidence

Prior to modeling, dependent variable ESTIMATED\_LOSS will be logged to achieve a more normal distribution and visualized to assess that data is a more normally distributed format.

A screenshot of a graph

Description automatically generated

Will load the caret package for its flexibility in model training, creation of dummy variables, preprocessing functions, and summary statistics of models. Will use a set.seed() function to ensure reproducibility. Will use dummyVars function for creation dummy variables of all categorical variables.

A computer screen shot of a code

Description automatically generated with medium confidence

After creating a data frame of dummy variables from chosen categorical variables, there is now a total number of 292 variables with 57,285 rows.

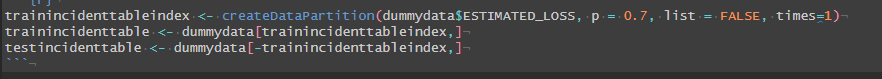


Due to limited CPU power available for such a large set of data, will be using a nearZeroVar() function. This will result in removal of variables that are associated with limited variance in the overall dataset. While it will likely reduce overall predictive performance of our models by a certain degree, it is necessary when working with limited computational resources.

A screenshot of a computer

Description automatically generated

To verify predictive performance of models, data will be partitioned into training and test sets. Models will be trained on training dataset, with model predictions evaluated on testing dataset. Split will be set at 70% training, 30% test. This is done in order to ensure our models are not overfitted to training data and can perform well on unseen, new test data.



To increase CPU capabilities for training large models, parallel processing will be enabled. See code below:

A screen shot of a computer program

Description automatically generated

For model training improvement, we will be using 5-fold cross validation. K-fold cross validation splits the data a certain number of times (k) and uses k-1 parts to train and the remaining part to test the data. This process repeats until all parts have been used for training and testing. This is performed in this analysis to prevent overfitting. As previously stated, our value for k will be 5.



Beginning training of multiple linear regression model, stepwise regression model, and random forest regression model. Preprocessing functions scale and center used to standardize the data distribution for our regression models. Scaling divides the data by the standard distribution and center ‘centers’ the data distribution around zero. This is performed to prevent variables that have a differing scale of values (Ex: 1-10 vs. 1-100) from having more influence on the model.







Using the models to make predictions on the testing dataset. Then drawing summary statistics from the results.

A screenshot of a computer program

Description automatically generated

Initial results from our dataset. Note that our dependent variable was logged so results will be indicative of that. Examining summary() results for our multiple linear regression model and our stepwise regression model. Then following with evaluation metrics of predictive performance on test dataset.

Summary() results for multiple linear regression model:

A black background with white text

Description automatically generated

Summary() results for stepwise linear regression model:

A black background with white text

Description automatically generated

Prediction results evaluation:

A screenshot of a computer program

Description automatically generated

With these results, the random forest model appears to offer the most predictive accuracy on the dataset. MAE “is the average absolute difference between the predicted and actual values” (Kumar, 2024). R Squared is an indicator of how much our model explains the variance in the dependent variable. “RMSE represents the square root of the variance of the residuals” (Bobbitt, 2021). The random forest model has the lowest MAE and RMSE, which indicates its predicted values are on average the closest to the actual values of the testing dataset. Its R squared value also indicates that it does a better job of explaining the variance in the dependent variable.

For reference, the statistics for logged ESTIMATED\_LOSS in the testing data is:



Now plotting the 25 most significant coefficients for multiple linear regression model (By p-value), stepwise regression model (By p-value), and random forest model.

A screen shot of a computer

Description automatically generated

A graph with blue dots and white text

Description automatically generated

A graph with a number of text

Description automatically generated

Obtaining P-values and coefficient estimates for multiple linear regression model and stepwise model. Along with attaching meaning of coded terms from NFIRS 5.0 reference guide to see which coefficients have the most significant p-values within the model, as well as determining their degree of positive or negative effect on ESTIMATED\_LOSS.

A screenshot of a computer program

Description automatically generated

For multiple linear regression model.

A screenshot of a computer

Description automatically generated



For stepwise regression:

A screenshot of a computer

Description automatically generated



Attaching residual plots between fitted values versus residuals for multiple linear regression model, stepwise regression model, and random forest model.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Random forest model residual plot.

A screen shot of a computer

Description automatically generated

While examining residual plots, it is evident that our models are not exhibiting Homoscedasticity. The residuals are not evenly distributed throughout and fitted values are not being predicted through the range of actual values.

**Data Summary and Implications**

The original research question was: Can a predictive regression model be constructed to estimate monetary loss from house fires that shows predictive improvement between multiple linear regression, stepwise regression, or random forest regression?

After review of our analysis and the successful creation of several predictive regression models: The null hypothesis can be rejected, and the alternate hypothesis is supported. Three different regression models were created, with both the multiple linear regression model and stepwise regression model showing statistically significant p-values (<0.05) in their respective summary statistics. Indicating that the models were statistically significant. Between our three models, the random forest did show improvement in all predictive accuracy metrics that were discussed earlier. Those metrics being MAE, R Squared, and RMSE.

While it appears evident that regression analysis can be used to predict the estimated monetary loss from house fires, there is certainly room for improvement in the existing regression models. The R squared values of all models show that overall, the models do a poor job of explaining the variance within the dependent variable. This is more apparent when examining the distribution of the residuals within residual plots. Fitted values appear to be clustered, can vary to a large degree, and are not spread equally throughout the range of actual test values. Another sign that the models need improvement in their predictive performance.

The limitation of current analysis is that predictor variables currently used do not entirely capture the extent of the dependent variable, ESTIMATED\_LOSS. Resulting in less-than-optimal predictive ability and lower R squared values. Although early regression models do indicate some predictive capability, improvement is needed. The course of action recommended is further experimentation and research into independent variables that can be used to improve predictive performance on the estimated loss from residential house fires.

In determining how to approach this, there are two possible approaches. If one wants to stay solely within the NFIRS dataset, higher computational power is needed. Earlier attempts at modeling this dataset prior to this analysis with a higher number of predictor variables and current workstation did yield regression models with higher evaluation metrics (MAE, RMSE). However, these early models resulted in computer crashes, data corruption/ deletion, or excessively long run times that rendered them not practical for exploratory analysis or review. A pivot to cloud computing resources or improved workstation capabilities will be needed to assist computational workload. In turn allowing model creation using a larger number of predictors to capture variance within the dependent variable.

If higher computational power via the cloud or hardware is not possible, another approach would be obtaining new predictor variables from external data sources and merging them into future models as another option. For example, NFIRS already has Zip code data for all incidents available. One could incorporate the median home value per zip code from an outside data source and merge it into existing NFIRS zip code data. It is likely many other such predictors on the costs of house fires could be found in external sources with further investigation.

In conclusion, three predictive regression models to estimate the costs of residential house fires were successfully created on the NFIRS dataset. In addition, the 25 most significant predictor variables for these models based on p-values, or the importance metric in the case of the random forest model, were found.

**Sources**

Insurance Information Institute. (n.d.). Facts + statistics: *Homeowners and renters insurance.* Insurance Information Institute. Retrieved October 5, 2024, from <https://www.iii.org/fact-statistic/facts-statistics-homeowners-and-renters-insurance>

U.S. Fire Administration. (2022, September 21). *Residential Fire Estimate Summaries*. U.S. Fire Administration. Retrieved October 21, 2024, from https://www.usfa.fema.gov/statistics/residential-fires/

*FAQ: Public Protection Classification*. (2015). Texas Department of Insurance. Retrieved October 22, 2024, from https://www.tdi.texas.gov/fire/fmppcfaq.html

U.S. Fire Administration. (2015) *Fire Incident Reporting System Complete Reference Guide*. Retrieved October 15, 2024, from https://www.usfa.fema.gov/downloads/pdf/nfirs/NFIRS\_Complete\_Reference\_Guide\_2015.pdf

IBM Cloud Team. (2024, August 21). Python vs. R: What’s the difference? IBM. Retrieved October 5, 2024, from [https://www.ibm.com/think/topics/python-vs-r#:~:text=R%20programming%20is%20better%20suited%20for%20statistical%20learning%2C,applications%2C%20especially%20for%20data%20analysis%20within%20web%20applications](https://www.ibm.com/think/topics/python-vs-r%23:~:text=R%20programming%20is%20better%20suited%20for%20statistical%20learning%2C,applications%2C%20especially%20for%20data%20analysis%20within%20web%20applications)

FEMA. (n.d.). ABOUT NFIRS. U.S. Fire Administration. Retrieved October 5, 2024, from <https://www.usfa.fema.gov/nfirs/about/#:~:text=After%20responding%20to%20an%20incident%2C%20a%20fire%20department,and%20civilian%20casualties%20and%20a%20property%20loss%20estimate>

FEMA. (2017, May). National Fire Estimation using NFIRS DATA. U.S. Fire Administration. Retrieved October 5, 2024, from https://www.usfa.fema.gov/downloads/pdf/statistics/national\_fire\_estimation\_using\_nfirs\_data.pdf

FEMA. (2011). *Fire Data Analysis Guidelines and Issues*. Retrieved on October 19, 2024, from [Review of *Fire Data Analysis Guidelines and Issues*].

Keboola. (2020, September 17). *The Ultimate Guide to Random Forest Regression*. Retrieved October 19, 2024, from https://www.keboola.com/blog/random-forest-regression

GeeksForGeeks. (2024, February 15). *What are the Advantages and Disadvantages of Random Forest?* Retrieved October 22, 2024, from https://www.geeksforgeeks.org/what-are-the-advantages-and-disadvantages-of-random-forest/

Shukla, S. (2017, December 1). *Regression in Machine Learning*. GeeksforGeeks. Retrieved October 20, 2024, from https://www.geeksforgeeks.org/regression-in-machine-learning/

Ravelo, C. (2022, September 16). *Stepwise Regression: What is it and should you use it? -*. Statistics Solutions. Retrieved October 24, 2024, from <https://www.statisticssolutions.com/stepwise-regression-what-is-it-and-should-you-use-it/>

Kumar, A. (2024, August 18). *MSE vs RMSE vs MAE vs MAPE vs R-Squared: When to Use?* Analytics Yogi. Retrieved October 25, 2024, from https://vitalflux.com/mse-vs-rmse-vs-mae-vs-mape-vs-r-squared-when-to-use/#Mean\_Absolute\_Error\_MAE

Bobbitt, Z. (2021, May 10). *How to Interpret Root Mean Square Error (RMSE)*. Statology. Retrieved October 26, 2024, from https://www.statology.org/how-to-interpret-rmse/

**NFIRS Dataset**

*Annual NFIRS Public Data | FEMA.gov*. (2023, June 2). Www.fema.gov. <https://www.fema.gov/about/openfema/data-sets/fema-usfa-nfirs-annual-data>