- 55908Machine Learning

Machine Learning Final Project

Tal Karaev & Doron Firman

Business Problem

Can insurance companies use data to increase their profitability?

The Datasets

We used three data sets:

- National fire program analysis system data - The main data set consisting of the majority of the fire information
- Climate data - A data set consisting of temperature and precipitation information for every month
- fips - A data set that contains the fips code translation for each state

The Datasets

The following are the main features from the complete data set:

FIRE_YEAR - The year the fire occurred

STAT_CAUSE_DESCR - The cause of the fire

FIRE_SIZE_CLASS - The class of the fire (A is the smallest, G is the largest)

STATE - The state in which the location occurred

DURATION_IN_DAYS - How many days the fire went on

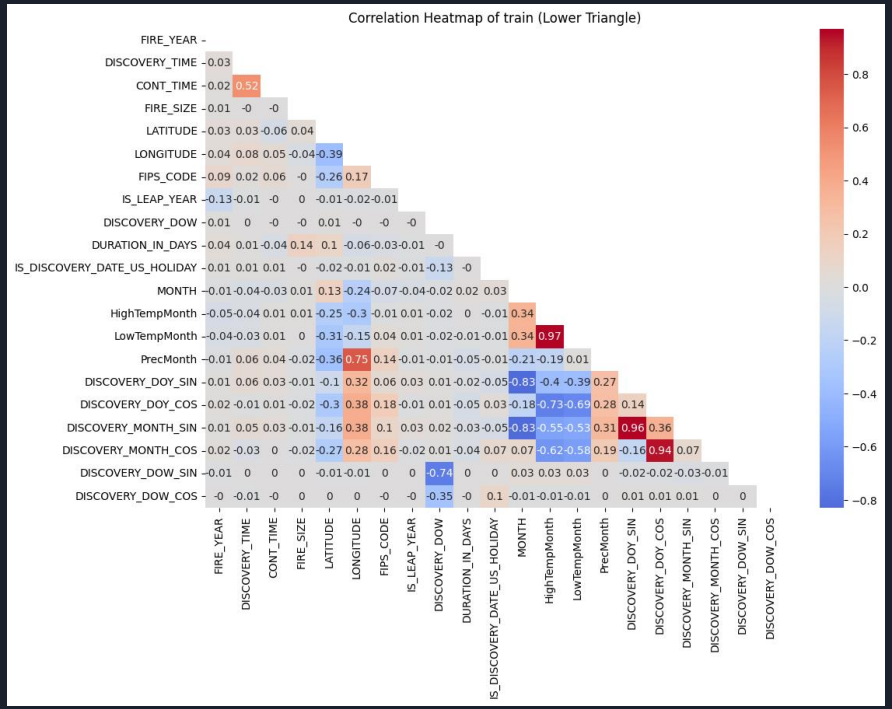
HighTempMonth - The highest temperature seen in the respective month

LowTempMonth - The lowest temperature seen in the respective month

PrecMonth - The average precipitation seen in the respective month

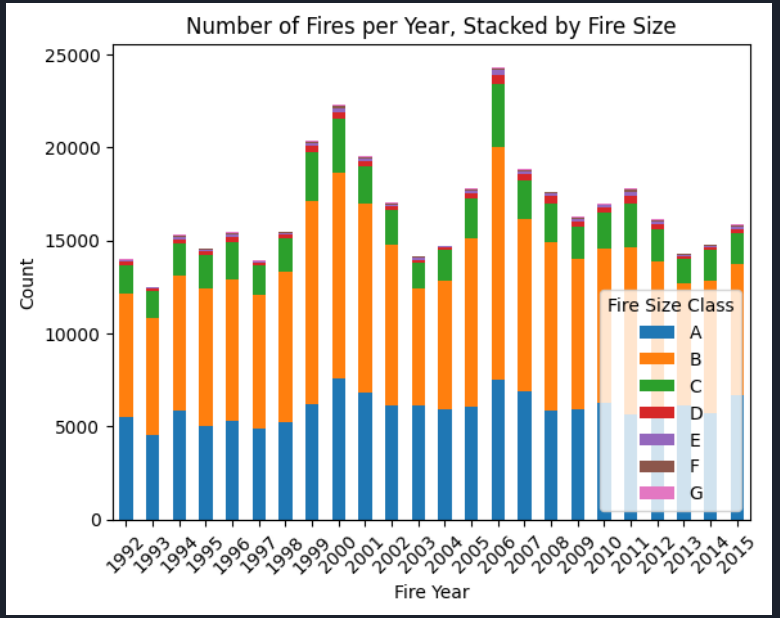
Exploratory Data Analysis

- As we can see in most of the matrix we don't have much correlation between variables
- We can see that we have higher correlation between features that describe time in different manners
- We also see a strong correlation between temperature features and to time features



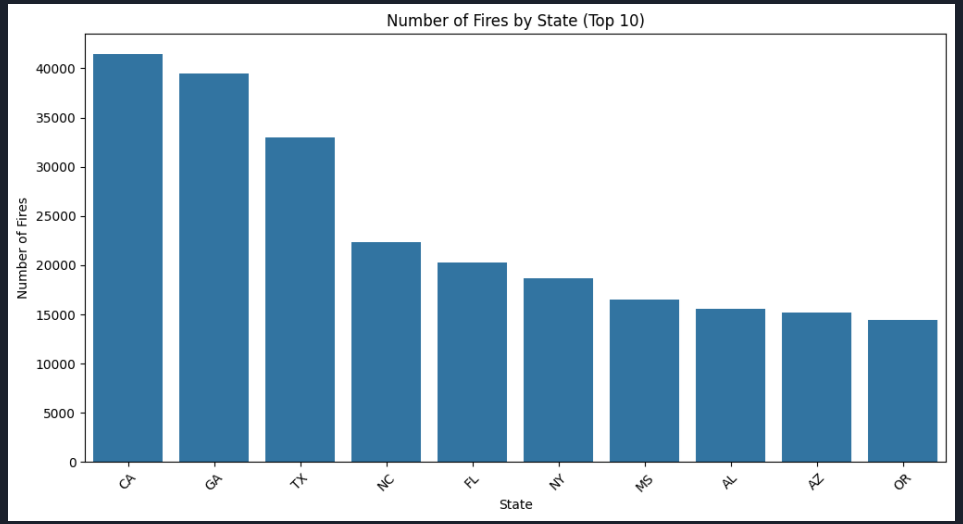
Frequency Analysis

- On average there are 10k fires each year
- Most of the fires are class B category fires
- Fire sizes of class F, G are very rare (as expected)



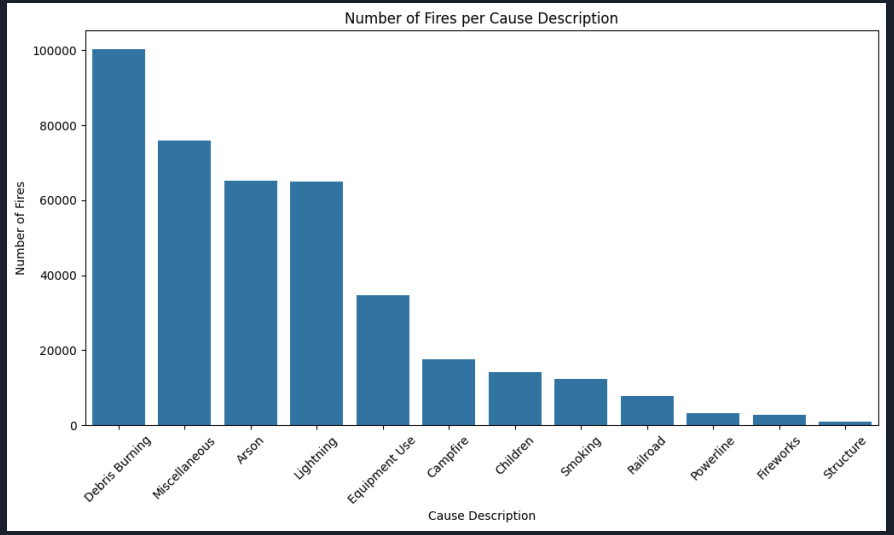
Frequency Analysis

- We can see that CA has the most fires historically
- We can see that there is a big difference between the first three states and the rest



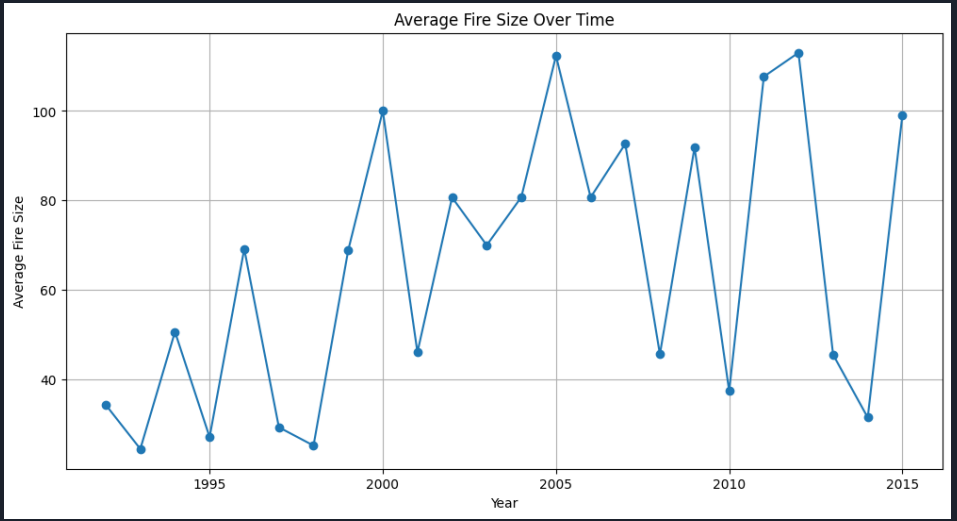
Frequency Analysis

- We can see that debris are most common
- We can see that there is a big difference between the first four causes and the rest
- Structure issues seems to be a very rare fire cause



Frequency Analysis

- We can see that the size of fires is rising



Models

We wanted to try and use the data to train a model that can predict, based on multiple fire parameters, what was the cause of the fire

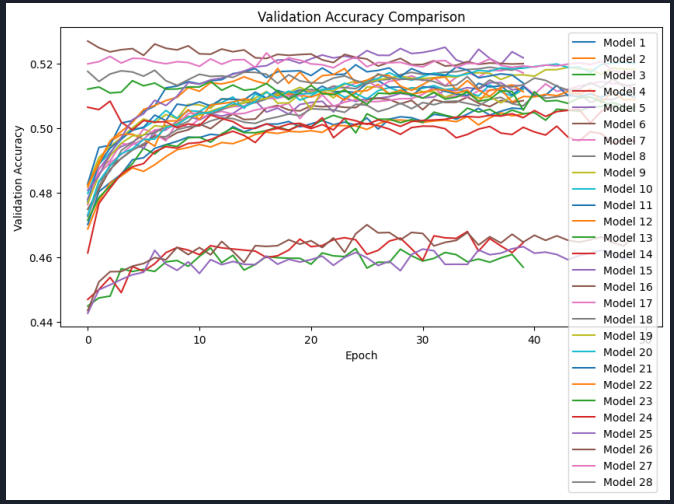
We created 7 different models which have different number and combinations of layers and types, this resulted in a total of 28 models

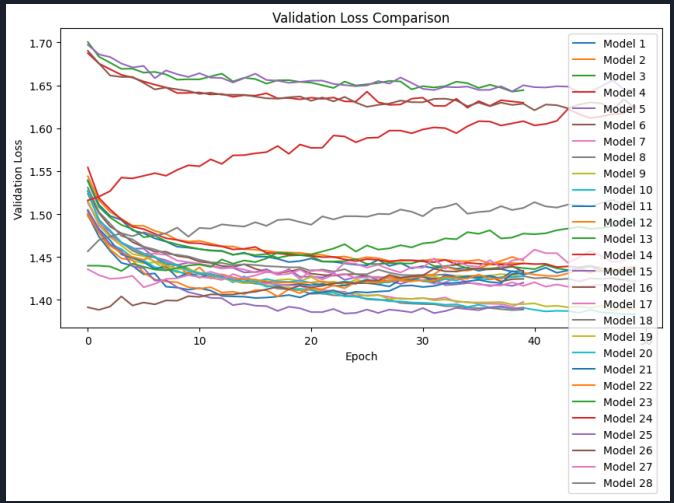
Hyperparameter Tuning

In order to receive the best possible model we ran each model in 4 different variations to find the best hyperparameters

We then compared all the models and picked and saw that the best three models were .25 ,27 ,26

They had the highest average validation accuracy and lowest average validation loss





Model Results (Train)

	Model 25	Model 26	Model 27
Average Validation Accuracy	0.516	0.522	0.52
Average Validation Loss	1.404	1.416	1.437
Combined Score	0.01	0.013	0.012

Model Results (Test)

Model 25

Test Loss: 1.406786

Test Accuracy: 0.518396

Model :26

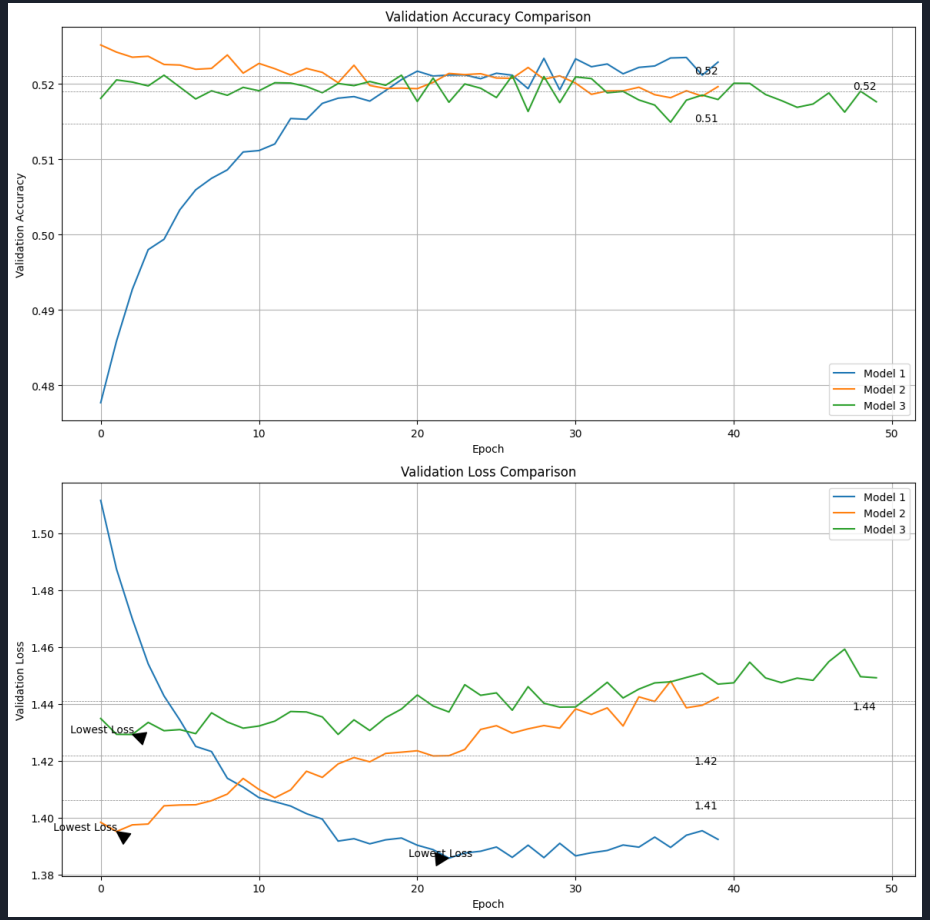
Test Loss: 1.463799

Test Accuracy: 0.512569

Model :27

Test Loss: 1.475661

Test Accuracy: 0.507476



Production Ready

In order to use the model out of the box we have made our top model production ready

- Download the zip file with the model pckl file and the script
- Add your own data set with the matching types
- Run the model

Conclusions

Using the data we gathered insurance companies can increase their profitability by:

- Reducing costs of insurance claims using our model that can help detect the cause of a fire based on its parameters
- Increased insurance priced based on location for states that have higher probability to have fires
- Increase revenue by creating insurance coverage packages that include less probable causes for fires

Future Research

Our best model reached a maximal validation accuracy of 58% which is not enough to successfully predict the cause of the fire in consecutive runs, we believe the following steps could improve the models performance:

- Our fire data is up until 2015 and more relevant data can assist with training the model and receiving higher results on today's cases
- Connecting an insurance company data set of fire damage claims can introduce new features which may highly impact our model performance
- As we saw increasing the amount of layers and neurons helped the model achieve better results so a more sophisticated model might achieve better results

Thank You