Advanced Analytics Final Paper – Spring 2020

Rodica Ceslov

Overview

This paper explores the application of machine learning techniques for predicting lethality in general aviation aircraft accidents and incidents. Machine learning is applied to a dataset derived from the Federal Aviation Administration (FAA) Aviation Accident and Incident Records database covering general aviation accidents in the United States from 1962 to 2019. The severity of injury was rated "Fatal" and "Non-Fatal".

The Dataset

I chose to limit my dataset to accidents that had occurred in the United States dataset because it has the largest general aviation community with hundreds of thousands of active aircraft according to the Federal Aviation Administration (FAA).

This is the process I followed and it summarizes the work I did on the dataset, for the Supervised and Unsupervised Learning Projects, and my takeaways.

Step 1: Data selection - Do I have the right data for the problem I am trying to solve?

I spent a significant amount of time trying to ensure that I have the right data for the problem I was looking to solve. I removed non-relevant columns. I then created a new column called "Lethality" which indicates *one* for Fatal accident and *zero* for Non-Fatal accident.

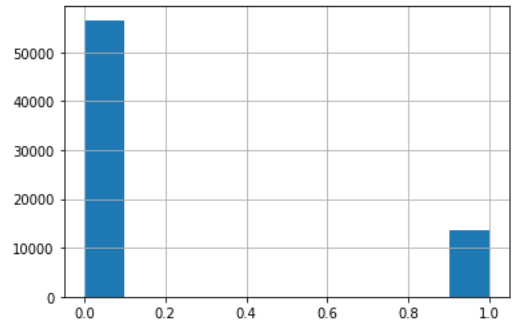
Assumptions:

I was careful about making assumptions about the dataset but there are several known facts which helped such as: FAA's data shows that most accidents are due to pilot error, that bad weather (IFR) is deadlier than good weather, and that small planes are much less safe than large commercial planes. I planned to test my assumptions and remove my own biases as much as I could, so I can be as objective as I can about the dataset.

Get a clear picture:
I explored the data through Jupyter notebook until I had a good picture about all the columns, rows and value counts. Plotting the data helped greatly in gaining a better picture. For example, I saw that the number of non-fatal accidents is far higher than the number of fatal accidents from the plot.

```
#Lethality
df_final["Lethality"].hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1f64
```



Data not available:
As I explored the data, I started to notice that there was data I wish I had. This was data that either did not get recorded or I did not have access to it. For example, I did not have details about the pilot such as flight hours, age, gender. Flight hours for each pilot is especially relevant to the kind of analysis I set out to do.
I also realized that in order to predict Lethality, it would be best to manually label it.

Data not needed:
There were columns that were either irrelevant of redundant but the decision of removing columns was not easy.

Step 2: Preprocessing data - Does the data have the right format and the features I need?

Formatting and cleaning:
This was more time consuming than anticipated. Although, the data was already formatted, it needed a lot of cleaning and removing or fixing missing data. A lot of data needed to be removed: helicopters, balloons, experimental aircraft, all aircraft data outside the United States. There were also numerous entries in an “Unknown” category. These are cases for which the cause is unknown or had not been determined. I decided to keep the unknown cases since removing them may give a flawed result.

Sampling:
Do I go beyond focusing on the dataset for United States accidents and use a smaller sample of the large dataset instead to ensure that running my models requires less memory?
I experimented with a smaller sample but ultimately I went with the entire United States dataset. After running the models though, I realized that I will need to rethink this.

Step 3: Exploring and transforming data – What transformations do I need to apply to the data to help ensure more accurate models?

As I went through the first steps, it became clear that the work done on the dataset was going to be critical to the accuracy of whatever model I choose and that this will be an iterative process.

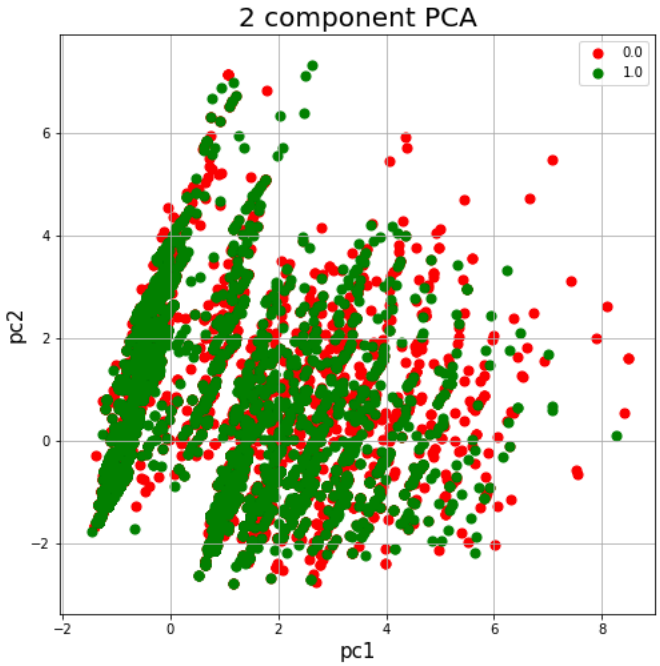
Scaling:
I tried scaling as that was one of the feature engineering transformations I recalled from class. This led me to labeling "Lethality" with *zero* and *one* as I stated previously. I saw this done on a dataset for the Titanic in a competition run by Kaggle.com. The idea was that I would do the same for airplane accidents: first record, fatal and not fatal, then explore them through the other data I have such as weather, make and model, manufacturer, latitude and longitude etc.

Outliers:
Plotted data to help detect outliers in the dataset.
Considered filtering out any outliers from the training set to test if model performance is impacted.
Used clustering methods such as the k-means to identify the natural clusters in the data recalling the example of the cluster centroids from class.
Applied Principal Component Analysis (PCA) as required in the assignment but it is unclear if helpful for this dataset.

– See Project 3 PCA plot.

```
finalDf = pd.concat([principalDf, train_df[["Lethality"]]], axis = 1)
```

```
fig = plt.figure(figsize = (8,8))
ax = fig.add_subplot(1,1,1)
ax.set_xlabel('pc1', fontsize = 15)
ax.set_ylabel('pc2', fontsize = 15)
ax.set_title('2 component PCA', fontsize = 20)
targets = [0.0,1.0]
colors = ['r', 'g', 'b']
for target, color in zip(targets,colors):
    indicesToKeep = finalDf['Lethality'] == target
    ax.scatter(finalDf.loc[indicesToKeep, 'pc1']
               , finalDf.loc[indicesToKeep, 'pc2']
               , c = color
               , s = 50)
ax.legend(targets)
ax.grid()
```



Binning:
I also looked at binning, grouping numerical value into bins but it didn't seem to make sense for my dataset.

Feature selection: What features to use for my predictive model?
I tried to go with the features that seemed the most useful but I learned that knowing the field would be needed for answering this question. However, it was unclear to me what feature selection and dimensionality reduction would do. I ran Principal Component Analysis (PCA) and I plan to spend more time understanding feature selection and dimensionality reduction.

Train and Test Datasets: Selected a train set and a test set and performed cross validation hoping to get a good balance between the train and test set. However, it is unclear what cross-validation accomplished on my dataset. This is a problem that I plan to spend more time on.

Testing Algorithms: I tested a few machine learning algorithms that were covered in class to see which performs best.

Models – Score Training + Test Set (Project 2)

	Model	ScoreTrainingSet	ScoeTestSet
3	Random Forest	84.09	84.13
8	Decision Tree	84.09	84.12
0	Support Vector Machines	82.42	82.64
4	Naive Bayes	82.38	82.53
2	Logistic Regression	82.23	82.35
6	Stochastic Gradient Decent	82.23	82.35
7	Linear SVC	82.23	82.35
5	Perceptron	80.72	81.00
1	KNN	80.03	80.34

Findings:
Running models led to issues with memory, crashing my machine. So working with smaller datasets may be better.
From the data exploration and plotting, it became clear that bad weather corelates with fatal accidents yet my models didn't show this correlation.

Takeaways:

1. The first takeaway from this process was that feature engineering would be key to my end result but it seems to be part art, part science and I had trouble figuring out how it would work on my dataset. I plan to spend time during the summer to learn more about it.
2. The second takeaway: I suspected that my dataset was imbalanced or there were other issues that I am yet to discover because the scores I was getting seemed misleading.
3. The third takeaway is that I don't know ahead of time which machine learning algorithm is the right one for my dataset but I do know if I have a supervised learning or an unsupervised learning problem.

Resources:
Data source is Kaggle.com: AviationData.csv. This is a dataset from the Federal Aviation Administration (FAA) Aviation Accident/Incident Database providing details about each aviation accidents from 1962 to 2019.
Titanic dataset source: Kaggle.com
AOPA: Aopa.com