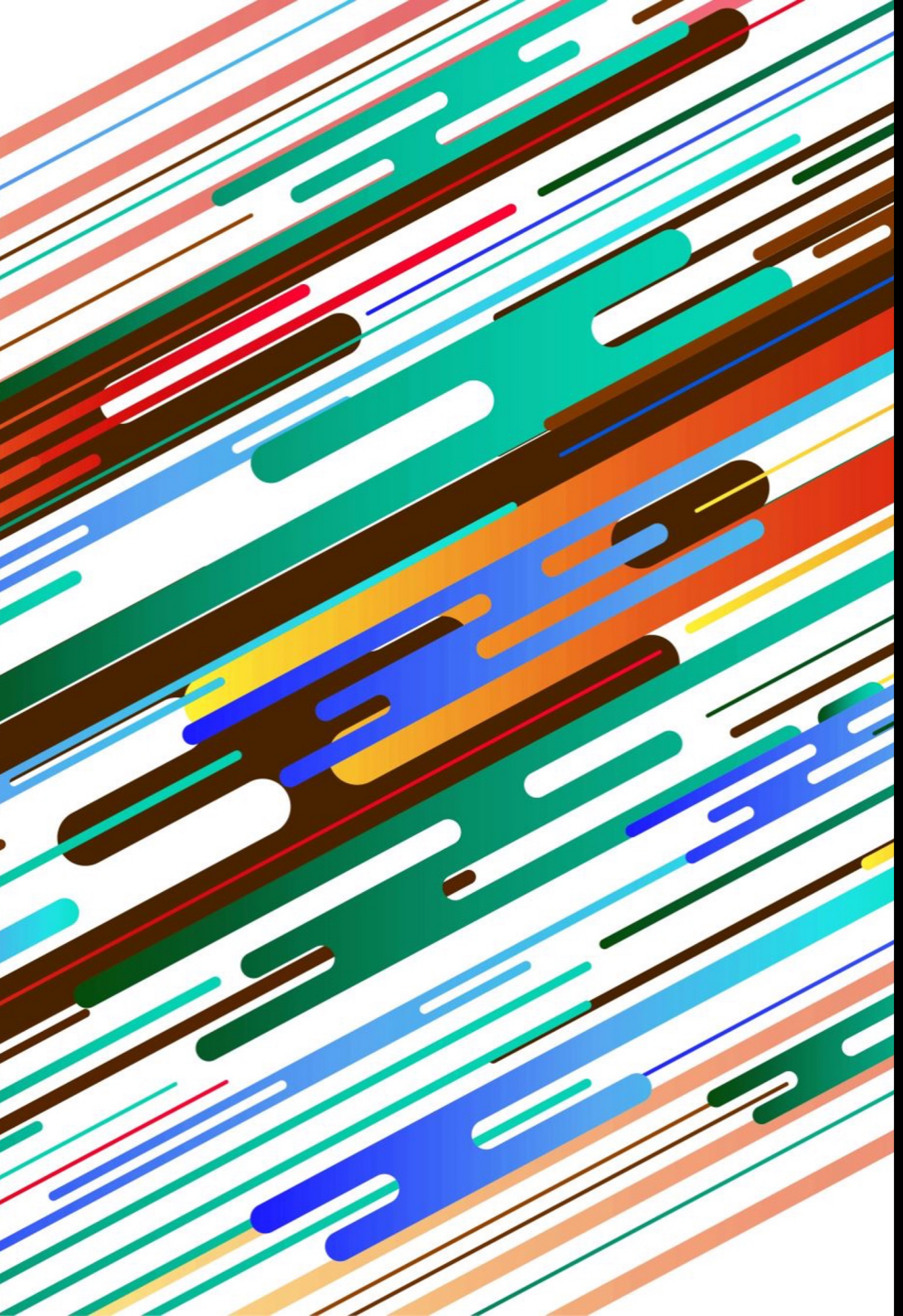


PROJECT #2 – MACHINE LEARNING APPLICATION – VEHICLE INSURANCE FRAUD CLAIMS

Brandon Allen

Mete Ozkazanc

Ryan Goda

INTRODUCTION & GOALS

We are a startup AI company.

We will be training a model to identify cases of vehicle insurance fraud and help predict cases of vehicle insurance fraud in future data.

Our goal is for our model to have an balanced accuracy of >75%.

Our Data was imported from Kaggle.

A Summary of our packages used include: NumPy, SkLearn, Pandas, and test_train_split.

OUR DATA & OBJECTIVES

We split our data into training and testing sets, applied the appropriate encoding, and tested different balancing techniques. Using the GradientBoostingClassifier and SMOTE Oversampling resulted in the best results.

BACKGROUND

Vehicle fraud is a deceptive practice to gain financial gain illegally

In the US, around 10%-20% of all insurance claims are fraudulent

AI can help impact the detection of fraud detection, pattern recognition & predictive modeling

COMMON FRAUDULENT CLAIMS

Staged Accidents

Double Dipping

Inflated Claims

Fake Injuries

Fronting

INITIAL FINDINGS

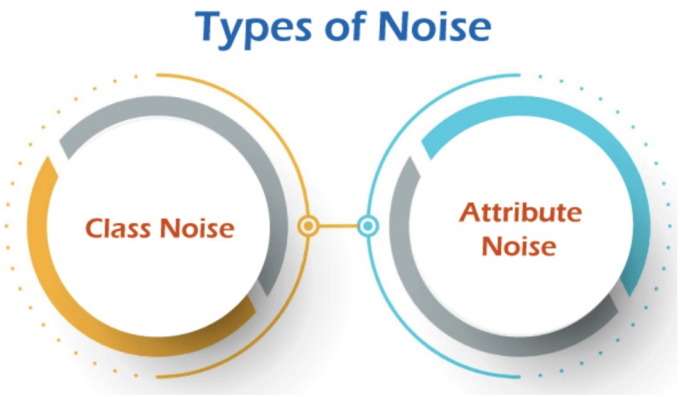
In building our Encoders, we initially attempted to utilize “DataMapper,” to help with our encoding. DataMapper has the ability to aid in encoding.

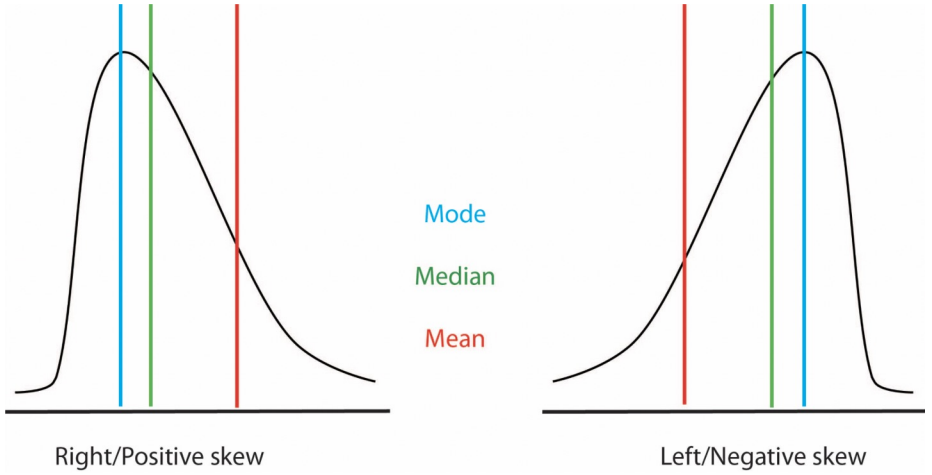
We found DataMapper to mangle the column names, leading to issues. Because of this, we built our own encoding pipeline.

CORRELATION MATRIX

We created a correlation matrix after our initial findings results.

The intent of the matrix was to identify features in our data that were “noise” to our training model, leading to skewed / inaccurate results.





LEARNINGS - TOP THREE

<class 'sklearn.ensemble._gb.GradientBoostingClassifier'> - SMOTE Oversampling

	precision	recall	f1-score	support
0	0.98	0.76	0.85	1451
1	0.15	0.69	0.25	91
accuracy			0.75	1542
macro avg	0.56	0.72	0.55	1542
weighted avg	0.93	0.75	0.82	1542

<class 'sklearn.tree._classes.DecisionTreeClassifier'> - SMOTE Oversampling

	precision	recall	f1-score	support
0	0.95	0.91	0.93	1451
1	0.17	0.29	0.21	91
accuracy			0.87	1542
macro avg	0.56	0.60	0.57	1542
weighted avg	0.91	0.87	0.89	1542

As seen above, our difference in F1 Score for 0 v. 1 is substantial. In short, the F1 score is a measure of the “Precision” and “Recall” evaluations.

<class 'sklearn.ensemble._gb.GradientBoostingClassifier'> - SMOTETomek Hybrid Balancing

	precision	recall	f1-score	support
0	0.97	0.79	0.87	1451
1	0.15	0.58	0.24	91
accuracy			0.78	1542
macro avg	0.56	0.69	0.55	1542
weighted avg	0.92	0.78	0.83	1542

Undersampling did not aid our results as it lowered the number of training rows.

DEMO

CONCLUSION

Due to the size of our dataset size being insufficient, we would like to further analyze the training of our model with a sufficient datasize.

In hindsight of our project, we would also like further analyze insurance data and see what applications our model has in other financial fields.

QUESTIONS?