Udacity Machine Learning Engineering Nanodegree: Capstone Proposal

Author: Matt Oehler

Domain Background

The project I will be working on entails working on a classification problem with severely imbalanced data. This is an issue that occurs frequently in practice, particularly in when building out anomaly detection systems or when trying to detect fraud. Early practices for dealing with imbalanced data include over-sampling from the minority class with replacement, or under-sampling randomly from the majority class such that it matches the size of the minority class. These methods while viable are not optimum solutions. Under-sampling from the majority class means that you aren't using all of the data that you have access to, while over-sampling means that you are using certain observations multiple times which may increase a model's propensity to overfit to certain behaviors in an undesirable manner. In this project I will be exploring to more modern approaches to dealing with imbalanced data. Both of these methods (SMOTE and ADASYN) are over-sampling methods that involve creating synthetic data based on the minority class. These solutions enable to users to both a) use all of the data they have access to and b) not duplicate any observations of the minority class. The methods will be discussed in more detail in the Capstone Report, but for the proposal I will simply refer to these to papers as academic resources for understanding the two methods in detail.

[ADASYN Paper]:
http:/scholar.google.com/scholar_url?url=https://sci2s.ugr.es/keel/pdf/algorithm/congreso/2008-He-ieee.pdf&hl=en&sa=X&scisig=AAGBfm1uJ1FIWcOyYTHBq1effUALxWAmHg&nossl=1&oi=scholarr

[SMOTE Paper]: https://arxiv.org/pdf/1106.1813

Problem Statement

The problem that this project is designed to solve is that of fraud detection. Fraud detection is difficult because fraudulent transactions typically make up a very small fraction of transactional datasets. The dataset we will be using is severely imbalanced with only 0.17% of the values being fraudulent. This severely hinders many of the most useful models in being able to detect a clear fraud signal and cleanly partition transactions as fraudulent or non-fraudulent. The ideal solution involves building a model that captures a vast majority of the fraud without blocking any of the non-fraudulent transactions (as this would lead to a very bad experience for legitimate customers). This is why precision and recall will be some of the primary metrics used to evaluate the models' performance. Potential solutions to this problem will explored using Logistic Regression and Tree Classifiers as models with the optimum solutions being determined from exploring different sampling strategies and parameter combinations. Details are expounded upon in following sections.

Dataset and Inputs

The dataset used in this project is publicly available and can be downloaded at the following link.

[Dataset]: https://drive.google.com/file/d/1CTAlmlREFRaEN3NoHHitewpqAtWS5cVQ/view

The dataset consists of over 280,000 transactions each with 30 features and a class label of fraud or not fraud. The class imbalance is very prevalent in this dataset as only 0.17% of the data points are labeled as fraud. Given that credit card transaction data includes personal identifiable information, the data provider anonymized the data by performing principal component analysis (PCA) on many of the features and removing most of the column names. All of the features are numeric in nature except for the class label which is simply a binary categorical feature. The feature list is specifically: Time, Amount, and V1-V28. Other details such as how time/amount were recorded or standardized are unavailable. The features (excluding the class label) will be explored and used to train the models assuming that

Solution Statement

I propose that we can effectively solve the issue of dealing with imbalanced transactional data through the use of cutting-edge methodologies, namely Adaptive Synthetic Sampling (ADASYN) and Synthetic Minority Over-sampling Technique (SMOTE). Research has shown that these newer sampling strategies can be very effective at overcoming the obstacles associated with class imbalance. Additionally, since this project will be evaluating models and results quantitatively, we should be able to determine how helpful these sampling strategies are, and perhaps, we will be able to confirm if one strategy is superior to the other.

Benchmark Model

In this project I will establish a few benchmarks to use as reference points. The primary benchmark will be a logistic regression model trained on data that has been balanced with random over-sampling. Both the model and sampling strategy are less cutting edge than the methods I will be experimenting with. As such, this model and sampling strategy will provide a great primary baseline. I also, for the sake of reference, created a 'model' that doesn't predict any fraud as well as a logistic regression model that is trained on the imbalanced data. These references will help to show the lift that data-balancing strategies bring to the table to bolster ML models in practice.

Evaluation Metrics

Accuracy is not a good metric when dealing with imbalanced data. The primary metrics I will use to evaluate the models in this experiment are metrics that can be derived from a model's confusion matrix. These metrics primarily include Precision and Recall. Given that we are dealing with fraud detection, I will place more weight on precision, as we'd like to minimize the number of false positives that the model produces as they result in a negative experience for the credit card user (e.g. declining legitimate transactions).

Project Design

Main objective:

- Build a model that can efficiently partition transactions into classes of fraud and not fraud such that fraud can be prevented without the potential detriment of good customer experience.

Process:

1. Explore data to ensure its integrity and make sure that all features are usable for modeling
2. Implement prepare the data such that it can be used to train various models
    a. This includes the process of using various sampling strategies to handle the issue of imbalanced data
3. Establish a few baseline metrics so that as a final solution is approached we have a method to quantify the usefulness of the proposed solution.
4. Iterate over various combinations of models and sampling strategies to find the top model(s)
5. Go through a process of hyperparameter tuning for the best performing model(s) to come up with the optimum solution.
6. Build out a web app with a user interface so that new transaction data can be fed into the model and yield results.