PROPOSAL

Insurance fraud

Kloeppel,Victor V.A.
Fontys hogeschool

Contents

Domain understanding

Domain understanding is an exploratory research this means that I am not going to discover new facts, but I am going to enhance my own understanding of the subject. For this research I am conduct an interview with the stakeholder.

What: For my project, I plan to use artificial intelligence to predict whether or not individuals are likely to commit fraud based on a dataset. Specifically, I will source a dataset of past instances of fraud and non-fraud cases, and use various machine learning algorithms to analyze the data and develop a predictive model. The goal of this project is to create a tool that can be used by organizations to prevent fraudulent activity and protect against financial losses.

Why: There are several reasons why I believe this project is important and worthwhile. Fraud is a serious problem that can have significant financial and legal consequences for individuals and organizations. Traditional methods of detecting fraud can be time-consuming and inefficient. AI-based approaches have the potential to automate much of the process and make fraud detection faster and more accurate.

Who: For my project I have an stakeholder that works as an accountant at a big firm. The stakeholder knows a lot about fraud and has found multiple cases of fraud at his work, he has useful information and tips that can help me make a model that can accurately detect fraud. For the project I am going to have an interview about his work and how to detect fraud. The stakeholder will also give me feedback on my project and will use the end result for testing.

When: The deadline of the project is the 7th of May. The proposal will be finished in the first few weeks so that I can begin as quickly as possible on my first iteration. The iteration zero will be done in week 7/8, from this point further iterations will be made to make as much improvements as possible before the deadline.

How: To complete my challenge I need a dataset, technology infrastructure for data processing algorithms. For the understanding I need the stakeholder to understand the data as good as possible. For the delivery phase I will work with the end users for testing.

What do I need to know about insurance to predict fraud?

How can you suspect someone of fraud without profiling them?

The goal is not to profile people, but research shows that certain groups are at lower risk of certain insurance fraud, so we will of course ask for less premium from those groups. Of course, we do not want to have a bias based on people's skin color or ethnicity..

How would this project be used in reality and what are the pros and cons?

I In reality, you cannot create a model that is 100% accurate in determining whether or not fraud has been committed. The model would be used as an initial analysis of an application, but it would not be a definitive conclusion of whether fraud has been committed. Later, an investigation will be started to further investigate the matter.

How can I make my system not unfairly target certain groups of people?

Of course, it is not the intention to group people and judge them based on that. We divide people into risk classes, and those with lower risk are given less attention. When it comes to data analysis, it is important to have non-discriminatory data beforehand, so ensure that the data used does not solely come from a certain origin.

How is fraud handled?

If fraud is suspected, the boss will be called first, and then a team of people with more work experience will go over the data again. We may never say definitively whether or not fraud has been committed, but we may only suspect it. Then a letter is sent asking for an explanation of why there is damage or why certain things are not in order with the application.

What are the points to look at when it comes to fraud?

Important points to consider for car insurance are how often damage has already been incurred. This also applies to other insurances, for example, people who regularly lose their phone while on vacation. When calculating a premium, we also look at your environment, where you live, how old you are, and whether you have had previous damage, which we see through claim-free years..

Data sourcing

The goal of this project is to build a machine learning model that can accurately detect fraudulent insurance claims. To achieve this goal, we need a dataset that contains information on both fraudulent and non-fraudulent claims, as well as relevant data points that can help us identify patterns and predictors of fraud. Some examples of these data point could be:

- Prices of the claim
- User id
- Time date of incident
- Insured interest
- Risk factor

To source data for this project, I explored various publicly available datasets on insurance claims. I identified a dataset on Kaggle that includes information on insurance claims, which I plan to use for our analysis. The dataset includes the following data points:

- User id
- Umbrella limit
- Policy start date
- Auto make
- Auto model
- Age
- Policy_csl

Analytic approach

- **Target variable**
  The target value is the feature you wish to predict based on the other selected features. For the project the target variable percentage of a chance that the request may be fraud

- **Algorithm selection**
  Different algorithms have different purposes, therefore the best algorithm should be picked particularly for this task. In the project it appears to be a classification and I need to look at what model would work best for this dataset

- **Good Indicators**
  For the ML model to make accurate predictions it is important that we have features that indicate characteristics that the model can use to link the link the fraud reports.