Project Proposal

Topic: Car Insurance Fraud
Team Member: Aniket Pandey, Manmeet Singh Khanna

Introduction: Fraud is one of the largest and most well-known problems that many insurance companies face. This project will focus on claim data of a car insurance company. Claims which are fraudulent can be highly expensive for companies. Therefore, it is important to know which claims are valid and which are not. It is not doable for insurance company to check all claims personally since this will cost too much time and money. In this project, we will take advantage of the largest asset which insurers have in the fight against fraud: Data. We will employ various attributes about the claims, insured cars and other circumstances which are included in the data by the insurer. Separating different groups of claims and the corresponding rates of fraud within those groups will provide us with new insights. Furthermore, we will be using Data Mining techniques to predict which claims are likely to be fraudulent. This information can narrow down the list of claims that need a further check. It will enable an insurer to easily detect fraudulent claims made by the policy holder. In addition, I, myself come from a finance industry background and know how important the insurance claims are for the banking and insurance sector. The major issue with fraudulent cases is that people who file a genuine case get delay in receiving their insurance money and a longer time is taken because of few cases where the fraud claims have been made.

Definition: The aim of this project will be to build a model that can detect car insurance fraud. The challenge behind fraud detection is that frauds are far less common as compared to legitimate insurance claims. It's a challenging problem, given the variation in fraud patterns and relatively small ratio of known frauds in typical samples. While building any detection or prediction models, the savings from loss prevention needs to be balanced with the cost of false alerts. Data Mining techniques will allow for improving predictive accuracy, enabling loss control units to achieve higher coverage with low false positive rates. Insurance frauds cover the range of improper activities which an individual may commit to achieve a favorable outcome from the insurance company. This could range from staging the incident, misrepresenting the situation including the relevant actors and the cause of incident and finally the extent of damage caused. If this problem is addressed this will tackle the above stated problem and bring clarity to claims being made by genuine customers.

Approach/Problems: The data available with us is from Kaggle which consists of 11564 rows and 35 columns. The data consists of following columns: accident area, accident date, details of the insurer, whose fault was it, vehicle type, witness present, policy report, age of vehicle, number of claims, etc. Although the dataset is small, we will find our way through this and bring something useful of whatever less or more we have. The ability to work with what is available is crucial for any company looking to transition into leveraging data science, compared to a company that waits for the day when it has a huge data set, the team that started with a small data set and worked on it will more likely succeed earlier in its data science journey and reap its rewards.
The next step is data visualization. It is one of the most important steps as we'll have to analyze the relation and decide on how the different columns are related to each other and what impact do they bring on the prediction. Before putting in the data through various predictive model we will first need to keenly visualize the data so as to get a clear picture what approach we are supposed to take and what will the outcome of our analysis. The data visualization and analysis will clear our picture around how we are supposed to use the data to bring a positive result to our aim of detecting fraud claims and preventing them from happening.
Model selection for the prediction is very crucial as the data is very sensitive of the feature. As a starter and study done from Literature surveys, we will employ the following algorithms: Random Forest, Isolation Forest, gradient boosting framework LightGBM, XGBoost, and SVM. We'll also implement different variations of the above algorithms and see which one performs best for our given data.

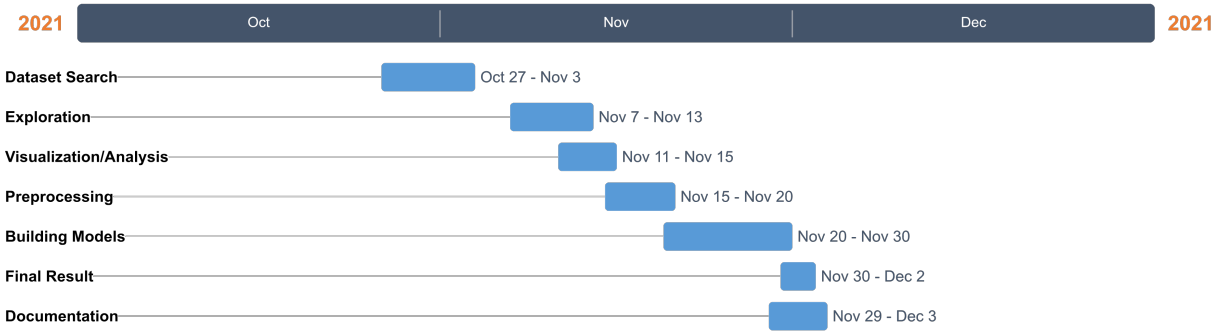
Literature Survey: There has already been many studies made on detection of fraud however there are very few dedicated to Car Insurance Fraud. Most of the research papers cover a general approach to the overall

insurance fraud. In my project I'll target a particular area of Car insurance and try to narrow down the research around it.

Phua et al. proposed a hybrid fraud detection model on a publicly available labelled dataset, which bring together the characteristics of stacking and bagging ensemble. Initially, the stacking classifier employs a meta classifier to select the best learner model from a bunch of base classifiers. The idea of applying fuzzy logic for identification of fraudulent claims among the settled cases has been suggested in Pathak et al.

A statistical bivariate probit model for analyzing and identifying false insurance cases has been developed by Pinquet. An asymmetric legit model by employing Bayesian method has been proposed in Bermúdez et al. for finding out illegitimate claims in a Spanish automobile insurance dataset.

Milestone: Here's a Gantt Chart with estimated days to represent timeline or the progress of the project



The progress of the Project can be tracked via github: aniket414/car-insurance-fraud. The repo will have latest and up to date code commits in master branch for tracking the progress.
We will employ the method of pair programming to work on this project. Github will help us for easy collaborations and completion of the task.

References:

G. Kowshalya and M. Nandhini, "Predicting Fraudulent Claims in Automobile Insurance," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 1338-1343, doi: 10.1109/ICICCT.2018.8473034.

Pathak, J., Vidyarthi, N., & Summers, S. L. (2005). A fuzzy-based algorithm for auditors to detect elements of fraud in settled insurance claims. Managerial Auditing Journal, 20(6), 632–644. doi:10.1108/02686900510606119

Phua, C., Alahakoon, D., & Lee, V. (2004). Minority report in fraud detection: classification of skewed data. ACM SIGKDD explorations newsletter, 6(1), 50-59.

Pinquet, J., Ayuso, M., & Guillén, M. (2007). Selection bias and auditing policies for insurance claims. The Journal of Risk and Insurance, 74(2), 425–440. https://doi.org/10.1111/j.1539-6975.2007.00219.x

C. Bernard, S. Vanduffel, Mean–variance optimal portfolios in the presence of a benchmark with applications to fraud detection. https://doi.org/10.1016/j.ejor.2013.06.023.