**Insurance Claim Fraud Detection and Premium Prediction using ML/Deep learning**

**DATA 698 Project Proposal**

**Jeyaraman Ramalingam**

**Background and Purpose**

**Insurance Fraud -** Insurance fraud is any act committed to defraud an insurance process. It occurs when a claimant attempts to obtain some benefit or advantage they are not entitled to, or when an insurer knowingly denies some benefit that is due.  Few examples of Insurance fraud include claiming for the benefits when the situation is not covered by insurance, increasing the impact of incident etc. According to FBI It is estimated that the average loss in United States due to Insurance fraud is 30 billion USD. Hence the insurance industry has urgent need to develop solutions which could eliminate the fraudulent claims with the state-of-the-art accuracy.

**Premium Pricing** - Another problem insurance companies face is striking a competitive premium pricing. An insurance price depends on various features such as age, type of coverage, amount of coverage needed, gender, body mass index (BMI), region, and other special factors like smoking to determine the price of the insurance.

**Disadvantages of Traditional Approach** – The fraud detection and pricing prediction is implemented by identifying certain parameters and the business professionals use mathematics and statistics for arriving at the numbers. This takes lot of time as the processes are manual and accuracy of the models are very less since all scenarios are not taken into consideration.

**ML/Deep Learning Approach** - Classification, Prediction and Artificial Neural Network models can be used to identify fraudulent claims and predict a premium price which could reduce potential losses to the insurance company.

**Research Questions**

The final set of questions will be identified upon investigating the datasets and through the model development process. Below are preliminary set of questions.

1. What are the factors can be used potentially to mark a claim as fraud?
2. What are the Customer demographics who are involving in Insurance frauds?
3. What are the common situations where the frauds are taking place?
4. How Anomalies in claims are compared to the reported frauds? Are all anomalies treated as frauds?
5. What are the factors impacting the premium price? What is the correlation between the response variable premium price vs the other variables?
6. What are the Customer demographics with different premium price ranges?
7. How can the parameters be altered to arrive at competitive price of insurance?

**Relevant Research**

There are quite a few research done by the large-scale organizations to improve the accuracy of fraud detection rate and the research available public on the fraud detection are on the lesser sample size datasets. Since there is an imbalance situation in the insurance datasets (less fraudulent claims comparing to the valid ones), sample size could play a key role in the accuracy standpoint. Also, most of the research are done using traditional ML algorithms (GLM). With respect to purchase premium prediction there are not many research which can be found online. There is not much research can be found online which combines the ML algorithms and deep learning algorithms.

**Research Methods**

There are not many real-world datasets which are publicized as this could create confidentiality issues. Out of few sample datasets are found on Kaggle, there is one fraud detection dataset prepared by Roshan Sharma has decent number of samples(n=1000) but not big enough to classify effectively. The dataset contains 1000 observations and 39 variables. The explanatory variables are several insurance metadata, and the response variable is fraud\_reported.

For the Premium price prediction, the idea is to build on the existing dataset from Kaggle which has the low sample size as well. The price prediction dataset contains 1338 observations and 7 features. The dataset contains 4 numerical features and 3 nominal features.

The plan is to leverage the datasets available on Kaggle and create more samples on top of it. This would provide a bigger sample size and accurate model results for both use cases.

The proposal is to build all necessary datasets and then perform exploratory data analysis on the dataset to find necessary correlations. Then build the classifier and regression algorithm models (Random Forest, Decision tree, SVM etc.) to sort the fraud/non fraud claims and implement Artificial neural network deep learning algorithm. Feature selection process will be used to include only the relevant variables in the model. Then compare the Accuracy of all models and implement Ensemble models to blend the already effective models to create a robust model for detection of fraud claims and predict the premium price. The expected output of this project is to efficiently sort and predict the fraudulent claims and predict the premium price.

**References**

[**https://towardsdatascience.com/for-real-auto-insurance-fraud-claim-detection-with-machine-learning-efcf957b38f3**](https://towardsdatascience.com/for-real-auto-insurance-fraud-claim-detection-with-machine-learning-efcf957b38f3)

[**https://www.plugandplaytechcenter.com/resources/detecting-insurance-fraud-machine-learning/**](https://www.plugandplaytechcenter.com/resources/detecting-insurance-fraud-machine-learning/)

**Fraud Detection Dataset -** [**https://www.kaggle.com/roshansharma/fraud-detection-in-insurance-claims**](https://www.kaggle.com/roshansharma/fraud-detection-in-insurance-claims)

**Premium Prediction Dataset -** [**https://www.kaggle.com/charanrajshetty/insurance-predictor-r2-score-0-85?select=insurance.csv**](https://www.kaggle.com/charanrajshetty/insurance-predictor-r2-score-0-85?select=insurance.csv)