Introduction

Insurance frauds cover the realm of undignified activities that an individual may commit in order to achieve a favorable outcome from the insurance company. It exploits an insurance contract. Insurance fraud is an illegal act on the part of either the seller or buyer of the insurance contract. From the issuer (seller) insurance fraud includes selling policies from fictional companies, failing to submit premiums and churning policies to create more commissions. Buyer fraud can include magnified claims, fake medical history, post-dated policies, viaticum fraud, faked death or kidnapping, and murder. We are going to handle the majority of the cases where the policyholder attempts to receive more money by overstating a claim.

Machine learning allows for improving predictive accuracy, enabling loss control units to achieve higher coverage with low false positive rates. In this paper, multiple machine learning techniques for fraud detection are presented and their performance on various datasets examined. The impact of feature engineering, feature selection and parameter tweaking are explored with the objective of achieving superior predictive performance.

Existence in Market

There are aspects around claim fraud due to which essentials of the insurance ecosystem are impacted in different degrees of severity. Following are some:

Underwriting: Claims Fraud impacts underwriting guidelines and policies and deteriorates the insurance risk pool.

Social costs: Due to claim fraud, the prices of insurance go up as a whole.

Encourages more frauds: When a successful fraud propensity of individuals to further indulge in fraud increases, thus encouraging more fraud.

Loss in reputation: Repetition of fraudulent claims for an insurance company causes loss in market reputation thus causing a decline in competitiveness.

Customer relationship: Fraudulent claims adversely impacts insurer's relationship with its existing customers and with prospects.

Loss of faith: Due to fraudulent claims people trust in insurance declines, which is detrimental to the growth of the insurance industry.

This shows the severity of insurance fraud. In Europe, fraud cases are mainly gang related, so the focus is typically on third party instead of first party fraud. Analytics engines may indicate where the fraud claims will shift to in the coming years, such as pet care division for some insurers. There is a huge auto fraud rise in US. The situation has worsened because of scarce resources, overworked staff, undetected fraud and changing behavior.
The insurance fraud detection is a discipline involving a set of analytical techniques that database to identify the vulnerabilities where fraud can happen.

T … billion by 2024 at a compound
a … lriving the growth of the industry
a

In India, when it comes to insurance frauds, rural India has taken the lead due to various reasons. Frauds are of different varieties and they mostly take place in rural and semi-urban areas where insurers do not have proper infrastructure to inspect or for that matter the local authorities, who are supposed to certify events, are corrupt.
Most frauds are by way of insurance taken on dead people. The moment arrests happened, claim intimation from Uttar Pradesh halved Manipulation was apparent when claims were not in proportion to distribution. Organized fraudsters identify people who are terminally ill, buy insurance on their behalf, and share the booty with the family members. There is a nexus between fraudsters, doctors, lawyers and village-level administrators.
In fact, it is estimated that the Indian insurance industry loses close to $6 billion to insurance fraud in India. This works out to about 8.5% of all the premiums collected every year.

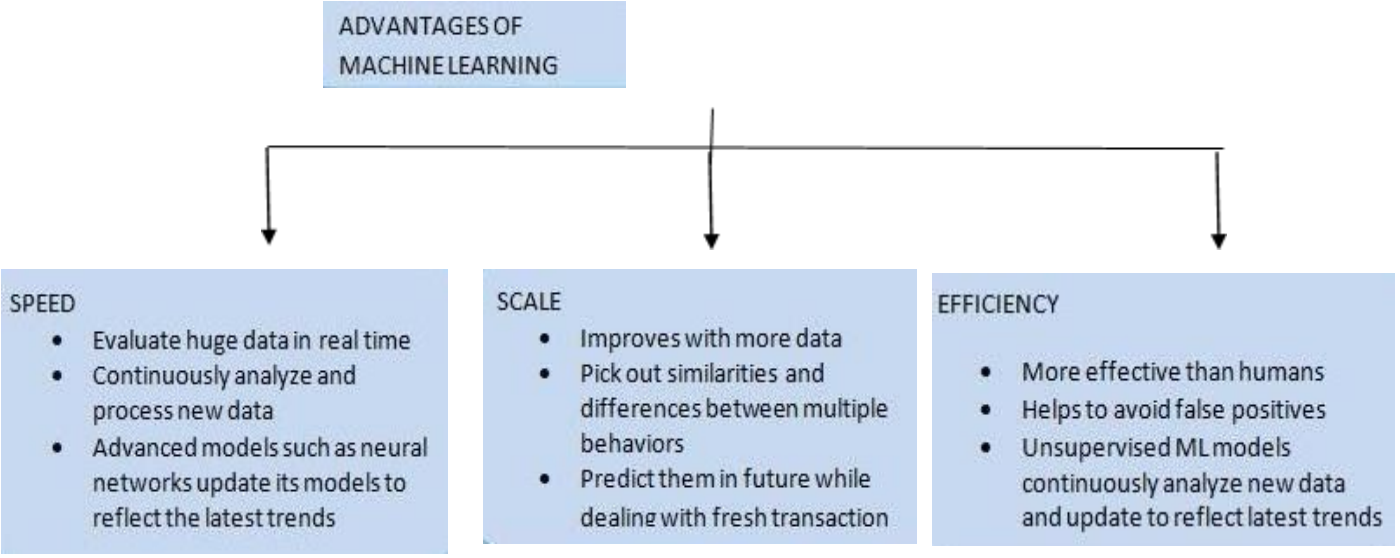
TECHNOLOGICAL ADVANCES IN THE INSURANCE INDUSTRY

With the advent of IoT, Machine Learning and usage-based coverage, big data all are making its way into the insurance industry. However, insurance companies still do not have a way to convert this data into meaningful insights. This is where Machine Learning help insurance companies structure their data in such a way that meaningful insights can be easily derived.

Why Machine Learning in Fraud Detection?

As stated by Fraud Benchmark Report by cyber source, 83% of North American business conduct manual reviews and on an average, they review 29% of orders manually. Participation of human gives intuition about fraud patterns and genuine customer behavior. However, manual review is expensive, time consuming and the greatest disadvantage is that it leads to high false negatives. This means completely normal customers just looking to make a purchase will go away from your business. A false positive not only affects the sale in the process but also lifetime value generated from the customer. Thus, manual reviews based on rules should be the last line of defense in the fraud detection strategy.

Machines are much better than humans are at processing large datasets. They are able to detect and recognize thousands of patterns instead of the few captured by creating rules. Prediction in large volume of data can be achieved by applying cognitive computing technologies to raw data.



Approach of detecting frauds in Insurance using Machine Learning:

1. **Data Cleaning:** Our dataset has 1000 rows and 39 columns. The columns with number of unique values in it are as follows:

```
months_as_customer 391
age 46
policy_number 1000
policy_bind_date 951
policy_state 3
policy_csl 3
policy_deductable 3
policy_annual_premium 991
umbrella_limit 11
insured_zip 995
insured_sex 2
insured_education_level 7
insured_occupation 14
insured_hobbies 20
insured_relationship 6
capital-gains 338
capital-loss 354
incident_date 60
incident_type 4
collision_type 4
incident_severity 4
authorities_contacted 5
incident_state 7
incident_city 7
incident_location 1000
incident_hour_of_the_day 24
number_of_vehicles_involved 4
property_damage 3
bodily_injuries 3
witnesses 4
police_report_available 3
total_claim_amount 763
injury_claim 638
property_claim 626
vehicle_claim 726
auto_make 14
auto_model 39
auto_year 21
fraud_reported 2
```

We dropped the columns which has maximum of unique values.

```
In [17]: for i in df:
             if '?' in df[i].unique():
                 print(i)

collision_type
property_damage
police_report_available
```

```
In [18]: print(df[(df["collision_type"]=='?') & (df["property_damage"]=='?') & (df["police_report_available"]=='?')].shape[0])
         for i in ['collision_type','property_damage','police_report_available']:
             print(df[df[i]=='?'].shape[0])

22
178
360
343
```

Since only 22 records had '?' in these 3 columns, we dropped 22 rows.
In the other rows they are replaced with not known.

We have generalized all the location and replaced them with ‘Drive’, ‘Ave’, ‘Ridge’, ‘St’, ‘Hwy’, ‘Lane’.

```
a=[]
for i,r in df.iterrows():
    if r['inc_loc'].split()[2]=="Drive":
        a.append('Drive')
    elif r['inc_loc'].split()[2]=="Ave":
        a.append('Ave')
    elif r['inc_loc'].split()[2]=="Ridge":
        a.append('Ridge')
    elif r['inc_loc'].split()[2]=="St":
        a.append('St')
    elif r['inc_loc'].split()[2]=="Hwy":
        a.append('Hwy')
    elif r['inc_loc'].split()[2]=="Lane":
        a.append('Lane')
print(set(a))
df['location']=a
df.location.unique()
```

```
{'Ridge', 'Ave', 'St', 'Hwy', 'Lane', 'Drive'}
```

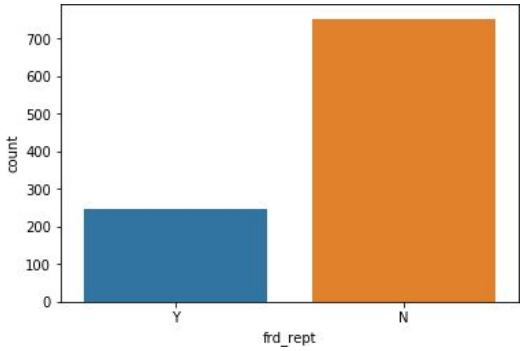
We manipulated the incident_date column and extracted the corresponding day of week, day of month and year and dropped the incident_date column.

2. **Data Analysis:** From several studies, following are some of the major indicators of Frauds in Insurance:

Policyholder Age: A closer look, however, reveals this attitude slightly more prevalent among younger policyholders than older ones. Similarly, a study published by the Insurance Fraud Bureau (2012) reveals that while 8% of all survey participants stated their willingness to participate in a staged accident for financial profit, this number increases to 14% among young people.
Hypothesis: H1: The younger the policyholders are, the more likely they are to engage in fraudulent activities.
Vehicle age: In connection with characteristics related to the insured vehicle itself, its age may be of interest to predicting the probability of a claim being fraudulent.
Older vehicles are more likely to be involved in fraudulent activities since policyholders may perceive its cash value as a form of additional funds when purchasing a new car.
Hypothesis: H2: The older vehicles are, the more likely they are to be involved in insurance claims fraud.
Vehicle type: Hypothesis: H3: The class of an insured vehicle has a significant impact on the probability of filing a fraudulent claim.
Vehicle value: It is composed of its catalog price and the value of any accessories, such as audio systems, car phones or air conditioning. In particular, these additions have the potential to substantially increase the insured vehicle’s value, the consequence being higher insurance premiums.
Hypothesis: H4: The higher the value of an insured vehicle, the more likely defrauding attempts become.
Type of damage: Particular focus was placed on loss events whose magnitude may easily be manipulated by either “overprovision” or “overcharging”. These include glass breakage and collisions.
Hypothesis: H8: Types of damages which are deemed to be difficult to verify (e.g., glass breakage and collisions) are more likely to contain fraud than those which are deemed easily verifiable.

The analysis from our data is as follows:

- It is an imbalanced dataset as very less number of fraud reported case are available.

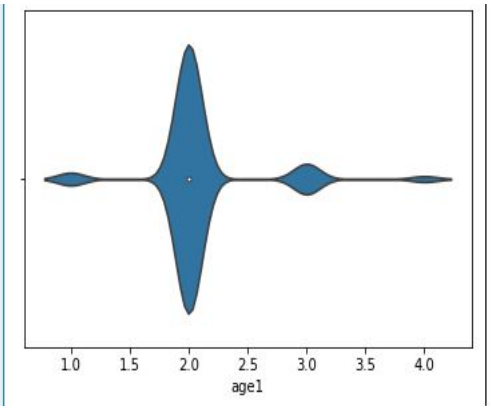


- People in age group 30-50 are mostly involved in fraud cases.

```
def clcol(m):
    if m<=25:
        return 1
    elif m<=50:
        return 2
    elif m<=60:
        return 3
    elif m>60:
        return 4
    else:
        return 5

df["age1"]=df["age"].apply(clcol)
for i in df["age1"].unique():
    print(i,df[(df.age1==i) & (df.frd_rept==1)].shape[0]/df["age1"].shape[0])

sns.violinplot(x="age",hue="frd_rept",data=df)
sns.boxplot(x="age", y="frd_rept",data=df)
```




```
In [94]: sns.violinplot(x="age",hue="frd_rept",data=df)
Out[94]: <matplotlib.axes._subplots.AxesSubplot at 0x2692ce
```

- If the total claim amount is greater than the total of injury claim, vehicle claim and property claim, then there is a higher chance of being fraud. In the dataset, no such trend was seen; it was equal in all cases. After applying feature importance, we dropped injury_claim, property_claim and total_claim_amount columns.

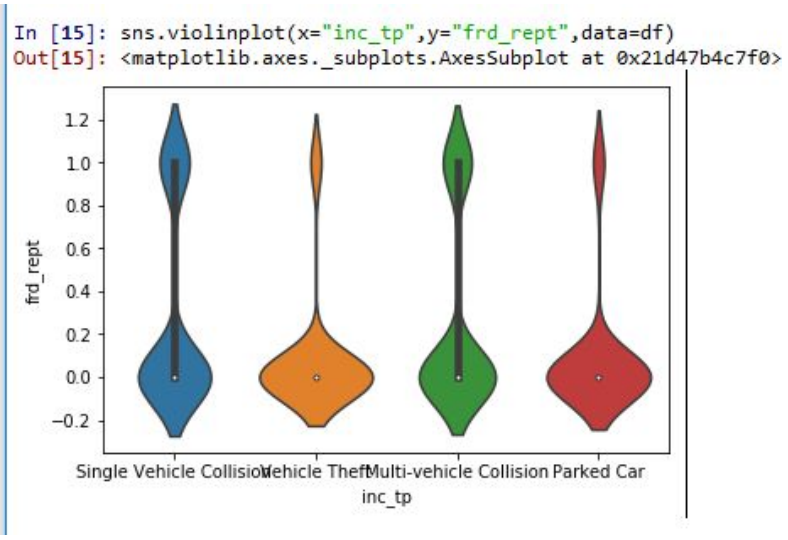
- People who have hobbies like playing chess, reading and crossfit, have more criminal bent of mind. Hence people who are involved in more brainstorming activities have more possibility of being involved in fraud cases.

- Highly educated people (MD, PhD and JD) and college students are more inclined in practicing fake claims.

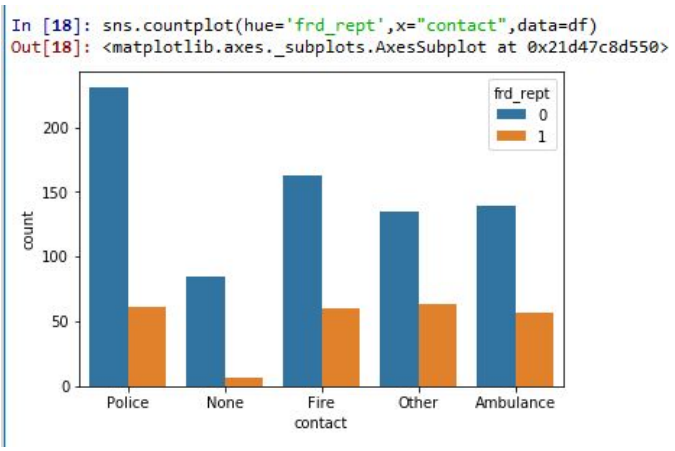
```
MD 0.2638888888888889
PhD 0.264
Associate 0.23448275862068965
Masters 0.22377622377622378
High School 0.225
College 0.26229508196721313
JD 0.2608695652173913
```

- People who own more expensive cars are more prone to fraud as they will have a higher claim amount.

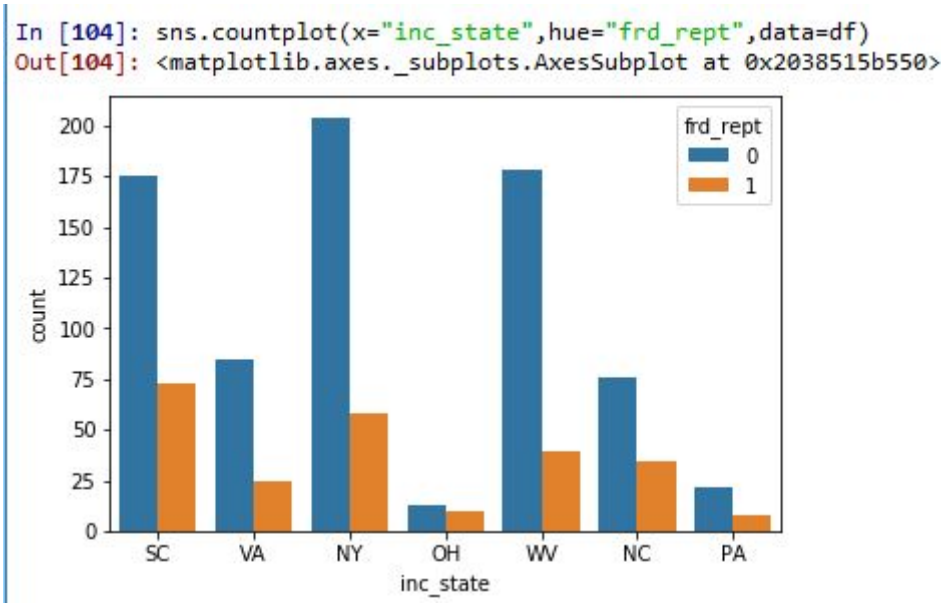
- Incidents of parked cars and car theft are less percentage in fraud cases as it is difficult to show that the car is lost. In collision cases, some parts are shown lost, which can be easily shown.



- People having contact with police, fire and ambulance have more chance of deceit insurance claims.



- NC and OH have higher chances of fraud cases.

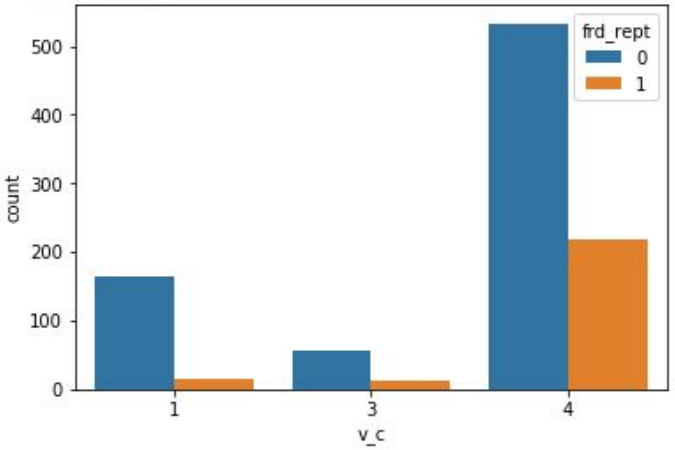


- Vehicle_claim with range (15000-30000) and (>30000) have more chances of getting fraud.

```
def clcol(m):
    if m<=15000:
        return 1
    elif m<=30000:
        return 3
    elif m>30000:
        return 4
    else:
        return 5

df["v_c"]=df["veh_claim"].apply(clcol)
for i in df["v_c"].unique():
    print(i,df[(df.v_c==i) & (df.frd_rept==1)].shape[0]/df["v_c"]
    [df.v_c==i].shape[0])
sns.countplot(x="v_c",hue="frd_rept",data=df)
```

```
Out[149]: <matplotlib.axes._subplots.AxesSubplot at 0x203855eac88>
```

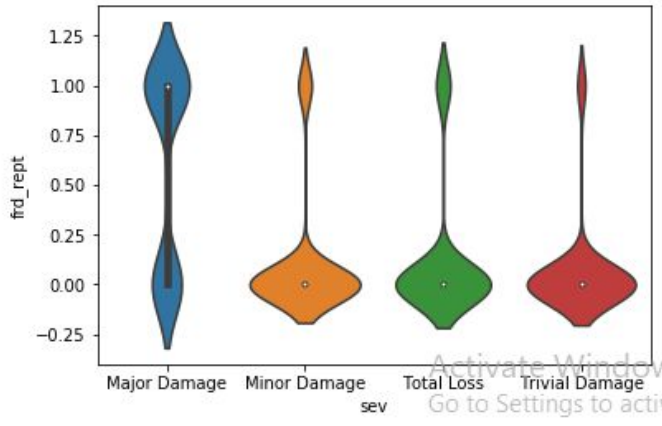



```
In [150]: for i in df["v_c"].unique():
     ...:     print(i,df[(df.v_c==i) & (df.frd_rept==1)].shape[0]/df["v_c"]
     ...:     [df.v_c==i].shape[0])
4 0.2912234042553192
1 0.08888888888888889
3 0.17647058823529413
```

- Major Damage severity has more chances of getting fraud.

```
In [106]: for i in df["sev"].unique():
     ...:     print(i,df[(df.sev==i) & (df.frd_rept==1)].shape[0]/df["sev"]
[df.sev==i].shape[0])
Major Damage 0.605072463768116
Minor Damage 0.10734463276836158
Total Loss 0.12857142857142856
Trivial Damage 0.06666666666666667

In [107]: sns.violinplot(x="sev",y="frd_rept",data=df)
Out[107]: <matplotlib.axes._subplots.AxesSubplot at 0x20385277048>
```



3. **Extract Data:** Here, the data will be split into three different segments – training, testing, and cross-validation. The algorithm will be trained on a partial set of data and parameters tweaked on a testing set.

4. **Building Models:** Building models is an essential step in predicting the fraud or anomaly in the data sets. We determine how to make that prediction based on previous examples of input and output data.

Many machine learning algorithms cannot operate on label data directly. They require all input variables and output variables to be numeric. So, for that we have used 2 methods:

1. One-Hot encoding: One hot (or dummy) coding for categorical features, produces one feature per category, each binary.
2. Label Encoding: Here, each unique category value is assigned an integer value. This is called a label encoding or an integer encoding and is easily reversible.

After this, various classification algorithms were applied like Logistic Regression, Decision Tree, Random Forest Algorithm, K-nearest neighbors and Naïve Bias algorithm.

```
encdf["hobb"]=df["hobb"]
X=encdf[["sev_0","sev_1","sev_2","sev_3","occ","edu_lvl","hobb","veh_claim","inc_tm","location","age1"]]
y=df["frd_rept"]
Xtrain,Xtest,ytrain,ytest=model_selection.train_test_split(X,y,test_size=.15,train_size=.75,random_state=42)
modelstats1(Xtrain,Xtest,ytrain,ytest)
```

```
{'n_neighbors': 5}
```

	MODELNAME	ACCURACY	PRECISION	RECALL	AUC
0	LR-train	0.758527	0.566667	0.093923	0.535186
1	LR-test	0.802721	0.600000	0.100000	0.541453
2	DecisionTree-train	1.000000	1.000000	1.000000	1.000000
3	DecisionTree-test	0.775510	0.451613	0.466667	0.660684
4	KNN(5)-train	0.776262	0.600000	0.281768	0.610087
5	KNN(5)-test	0.693878	0.173913	0.133333	0.485470
6	NB-train	0.803547	0.610778	0.563536	0.722891
7	NB-test	0.795918	0.500000	0.566667	0.710684

The best results came from Naïve Bias Algorithm on applying it on incident_severity, occupation, education_level, hobbies, vehicle_claim, incident_time, location, age with following value:

	TRAIN DATA SET	TEST DATA SET
ACCURACY:	0.803547	0.795918
PRECISION:	0.610778	0.50000
RECALL:	0.563536	0.566667
AUC:	0.722891	0.710684

CONCLUSION AND FUTURE WORKS

Although it is known that there is always a chance for betterment, this is the best model we could get. And it was achieved using Naïve Bias classification algorithms. It can be applied in the insurance ecosystem to make predictions in real time, and detect such fraud with more accurate and précised results.

In future, there is enough room for improvement in our project like:

1. Implementation of boosting techniques (adaboost, xgboost)
2. Implementation of SVM, neural networks
3. Using the columns policy_csl, umbrella_limit, policy_bind_date and months _as_customer could yield better results

But due to lack of time and our knowledge we could not implement them. Since research is an endless process, in future we can improve it by having a better analysis on the dataset and implementing new technologies.