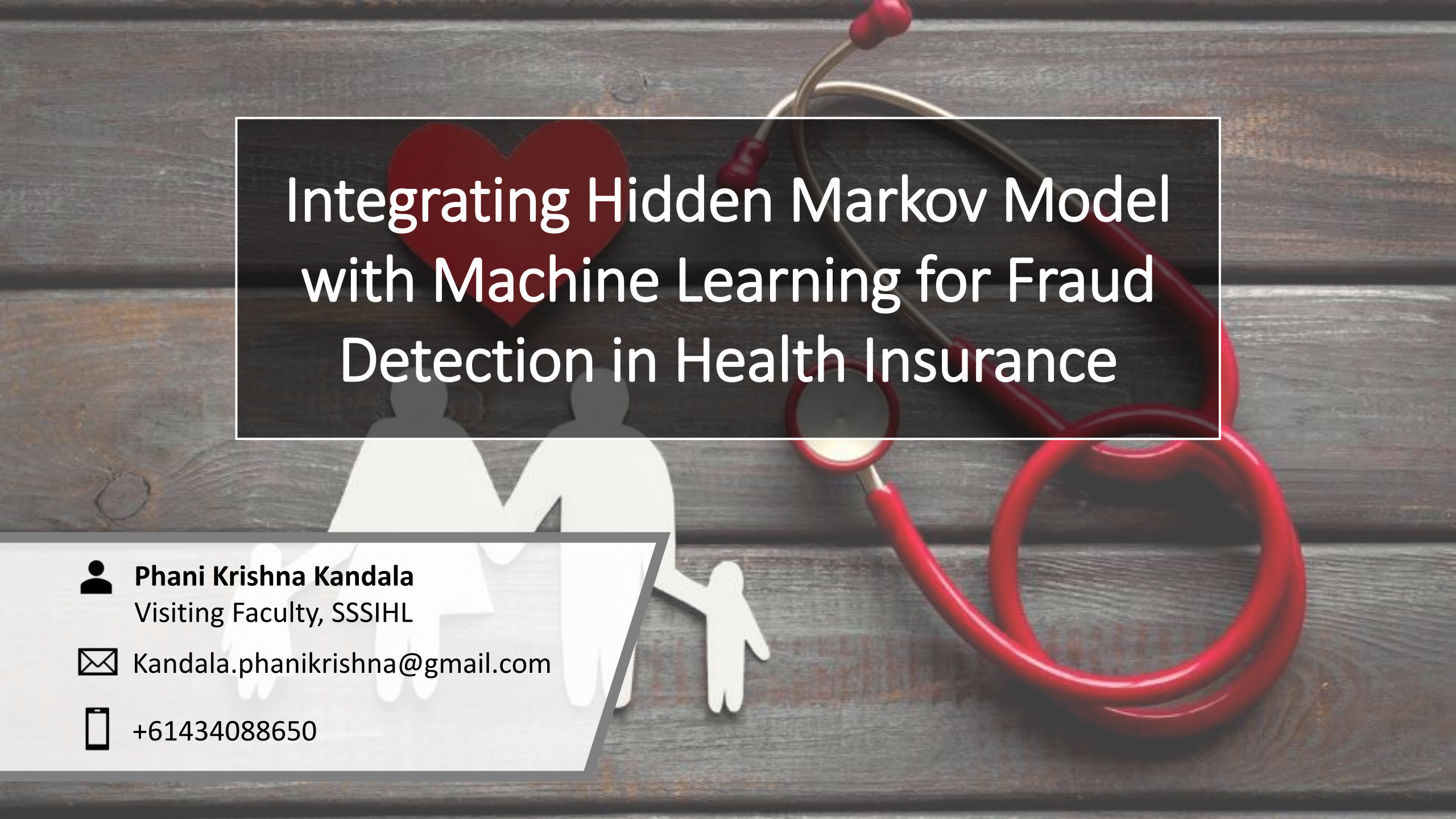


Integrating Hidden Markov Model with Machine Learning for Fraud Detection in Health Insurance

Phani Krishna Kandala
Visiting Faculty, SSSIHL

Kandala.phanikrishna@gmail.com

+61434088650

Agenda

Impact

Dataset

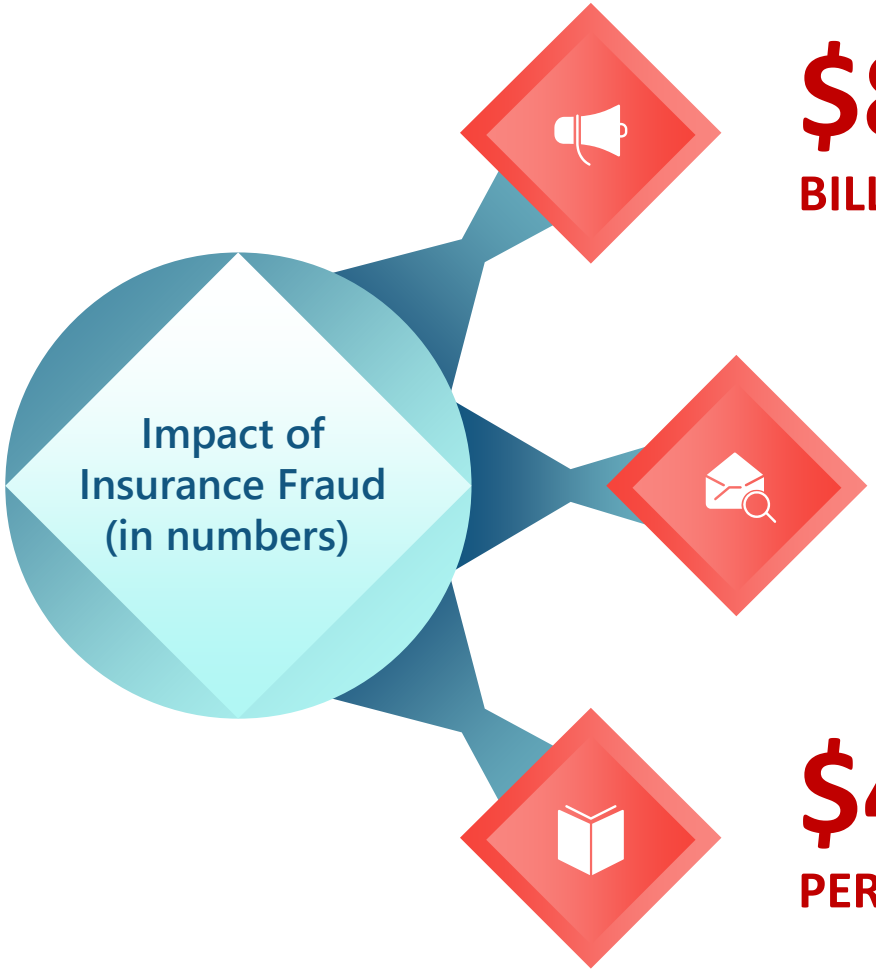
Proposed System

Modelling Details

Results

Conclusion

Impact



$80
BILLION

Costs to U.S consumers due to Insurance fraud

$40
BILLION

Non-health insurance portion of the overall

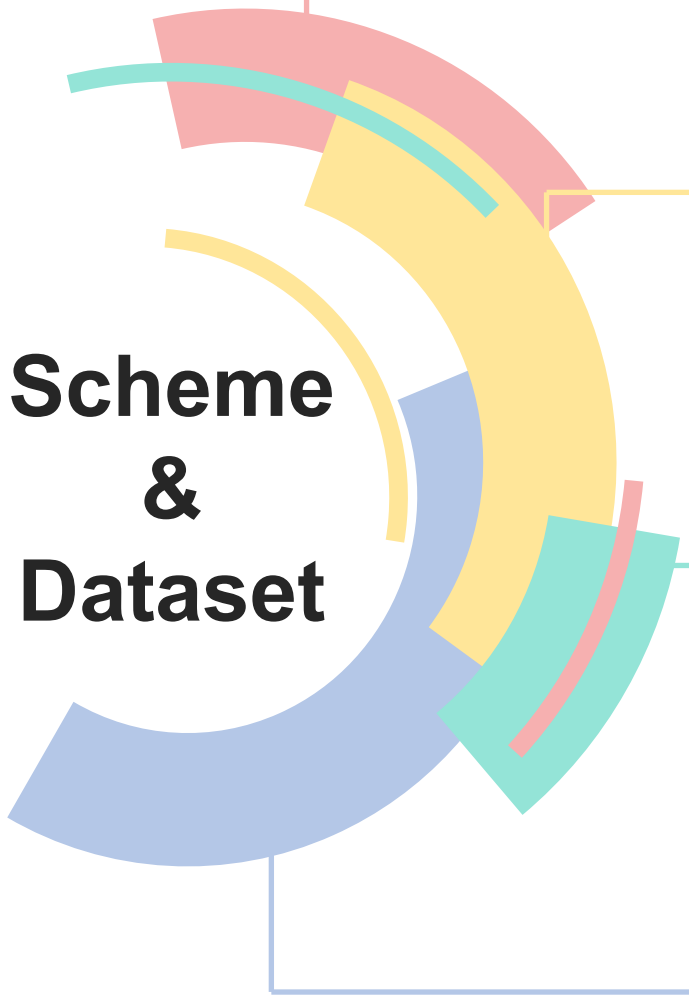
$400 to $700
PER FAMILY

Amount that average U.S. family will pay extra per year in premiums

https://www.iii.org/article/background-on-insurance-fraud#

Universal Group Health Insurance Scheme

World's largest



Scheme & Dataset

SCHEME

-

500+ million covered
2500+ packages
24000+ hospitals empaneled

TIME PERIOD

-

Aug-2019 to Aug-2020

DATA SIZE

-

380,000+ records
26 features

FRAUD CASES

-

~ 12% fraudulent cases

Markov Models

Set of states:

- $\{s_1, s_2, \ldots, s_N\}$

Process moves from one state to another generating a sequence of states:

- $s_{i1}, s_{i2}, \ldots, s_{ik}, \ldots$

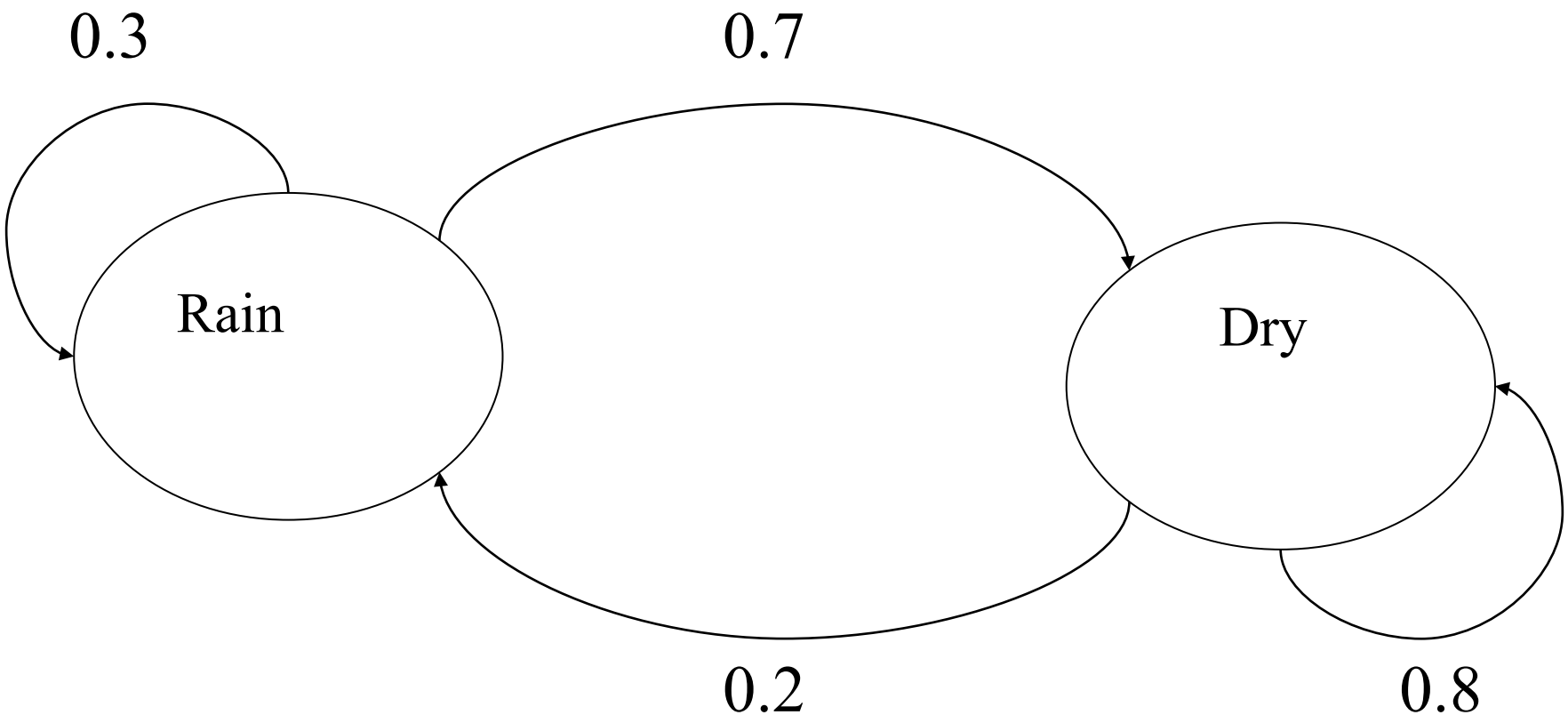
Markov chain property: probability of each subsequent state depends only on what was the previous state:

- $P(s_{ik}|s_{i1}, s_{i2}, \ldots, s_{ik-1}) = P(s_{ik}|s_{ik-1})$

To define Markov model, the following probabilities have to be specified:

- Transition probabilities $a_{ij} = P(s_i|s_j)$ and
- Initial probabilities $\pi_i = P(s_i)$

Example of Markov Model



Calculation of sequence probability

By Markov chain property, probability of state sequence can be found by the formula:

- $P(s_{i1}, s_{i2}, \dots, s_{ik}) = P(s_{ik}|s_{i1}, s_{i2}, \dots, s_{ik-1})P(s_{i1}, s_{i2}, \dots, s_{ik-1})$
  $= P(s_{ik}|s_{ik-1})P(s_{i1}, s_{i2}, \dots, s_{ik-1}) = \cdots$
  $= P(s_{ik}|s_{ik-1})P(s_{ik-1}|s_{ik-2}) \dots P(s_{i2}|s_{i1})P(s_{i1})$

Suppose we want to calculate a probability of a sequence of states in our example, {'Dry','Dry','Rain',Rain'}.

- P({'Dry','Dry','Rain',Rain'})
  - = P('Rain'|'Rain') P('Rain'|'Dry') P('Dry'|'Dry') P('Dry')
  - = 0.3*0.2*0.8*0.6

Hidden Markov models

Set of states:

- $\{s_1, s_2, \dots, s_N\}$

Process moves from one state to another generating a sequence of states:

- $s_{i1}, s_{i2}, \dots, s_{ik}, \dots$

Markov chain property: probability of each subsequent state depends only on what was the previous state:

- $P(s_{ik}|s_{i1}, s_{i2}, \dots, s_{ik-1}) = P(s_{ik}|s_{ik-1})$

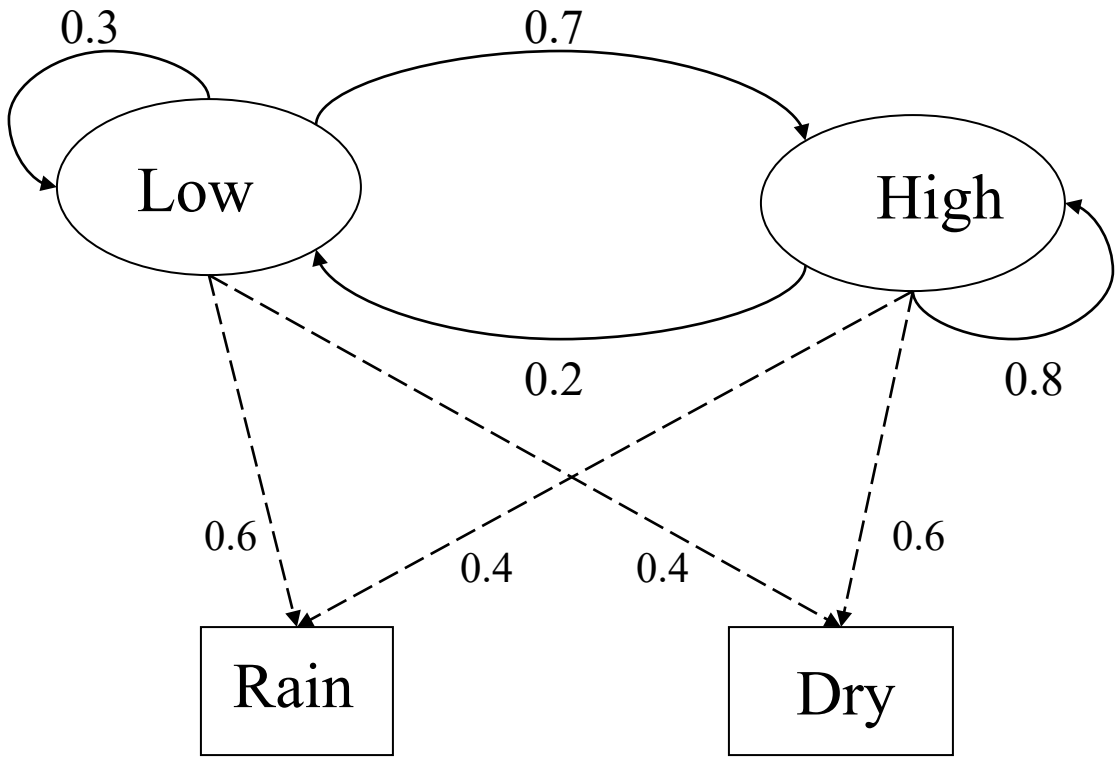
States are not visible, but each state randomly generates one of M observations (or visible states) $\{v_1, v_2, \dots, v_M\}$

To define hidden Markov model, the following probabilities have to be specified:

- matrix of transition probabilities $A=(a_{ij})$, $a_{ij}= P(s_i \mid s_j)$,
- matrix of observation probabilities $B=(b_i (v_m))$, $b_i(v_m) = P(v_m \mid s_i)$; and
- vector of initial probabilities $\pi=(\pi_i)$, $\pi_i = P(s_i)$

Model is represented by $M=(A, B, \pi)$.

Example of Hidden Markov Model



Fraud Detection using HMM

Using HMM for fraud detection is advantageous in two ways:

- Only outcome is visible to external observation environment and not the states.
- Although not every case, but most often, we see a reduction in false positives.

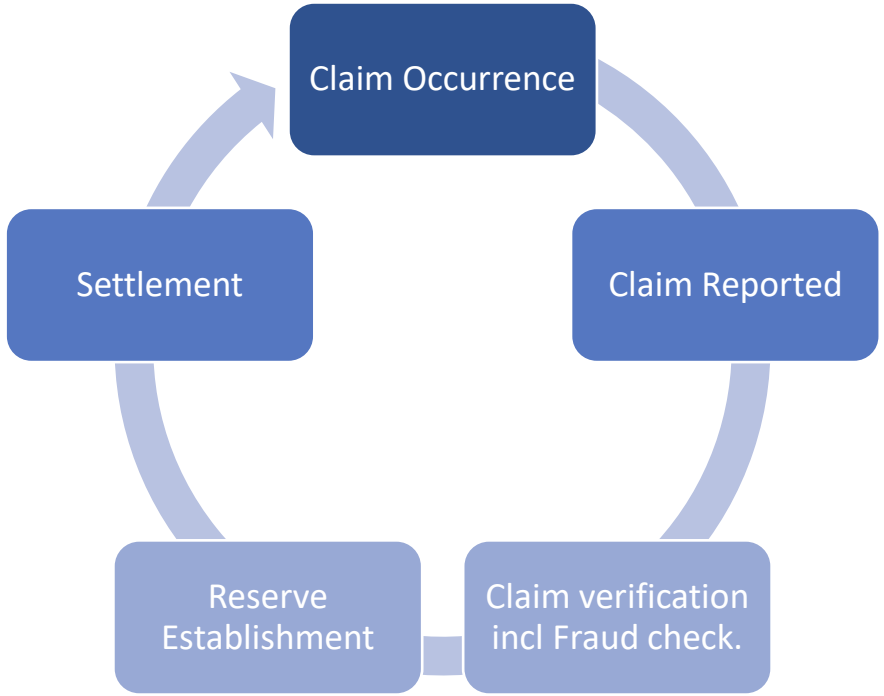
System we are working on does require the claiming behaviour, mainly utilisation of benefit packages.

Generally, the issue is, insurer is not aware of the granular diagnosis details either during the treatment or post treatment.

If we go through the Claims cycle, fraud detection comes under the claim's verification segment

Prediction process for HMM would consist of three categories mainly:

- Low utilisation
- Medium Utilisation
- High utilisation



Fraud Detection using HMM … Cont

Each individual claim does go through the fraud detection system for verification purpose. The fraud detection system takes the claim details such as nature of treatment, patient's age, previous health conditions, type of package, the utilization of package to validate, whether the claim is genuine or not.

Using HMM, in order to detect a fraudulent claim, it creates training set clusters and identifies the claiming behaviour of the policyholders.

This work mainly concentrates on amount of package benefit each claimant utilised and would classify into three categories mentioned earlier.

Operationally, it stores data of different claim amounts in form of clusters depending on utilization of the package benefit which will be either in low, medium or high utilisation range.

Then the model tries to find the variance in the claim amounts depending on each individual claim amounts.

Claims data would bring in additional information in the form of distinct characteristics to build the claims profile for the portfolio.

Algorithm

Bucketize the individual claim amounts into the reasonable ranges such as x1, x2, …,xM

Variance: Genuine differences in the claim size are captured based on individual claim circumstances.

Clustering: HMM determines the claim amount range using clustering algorithm say k-means.

Prediction process: would summarise into three claim cost ranging from low to high utilisation costs.

Probability Establishment: Uses the deviation established in different individual claim amounts say latest 100 claims to obtain the sequence and then establish probability

Initial value determination: Initial values are obtained from Market/expertise/subjective judgement.

Additional Complexity: In the recent times, usage of online claim filing does not detect fraud. For example: ransomware attack or phishing attack might make it more difficult to trace fraud.

How Fraud detection works

Claims and policy data is fed into the detection system to ensure the initial verification of contract such as contract expiry date, terms and conditions etc are checked.

If there is inadequate claims data, then exposure detection using individual claim related details are used. Else claims data is utilised to perform experience detection based analysis.

By using claims data we develop relationship to the observation by forming intermediate states leading to output observed states.

Transition probabilistic calculation of the states based on HMM are generated.

Conditional Probabilities obtained above in step 4 would give us an indication of a claim being fraud or not fraud.

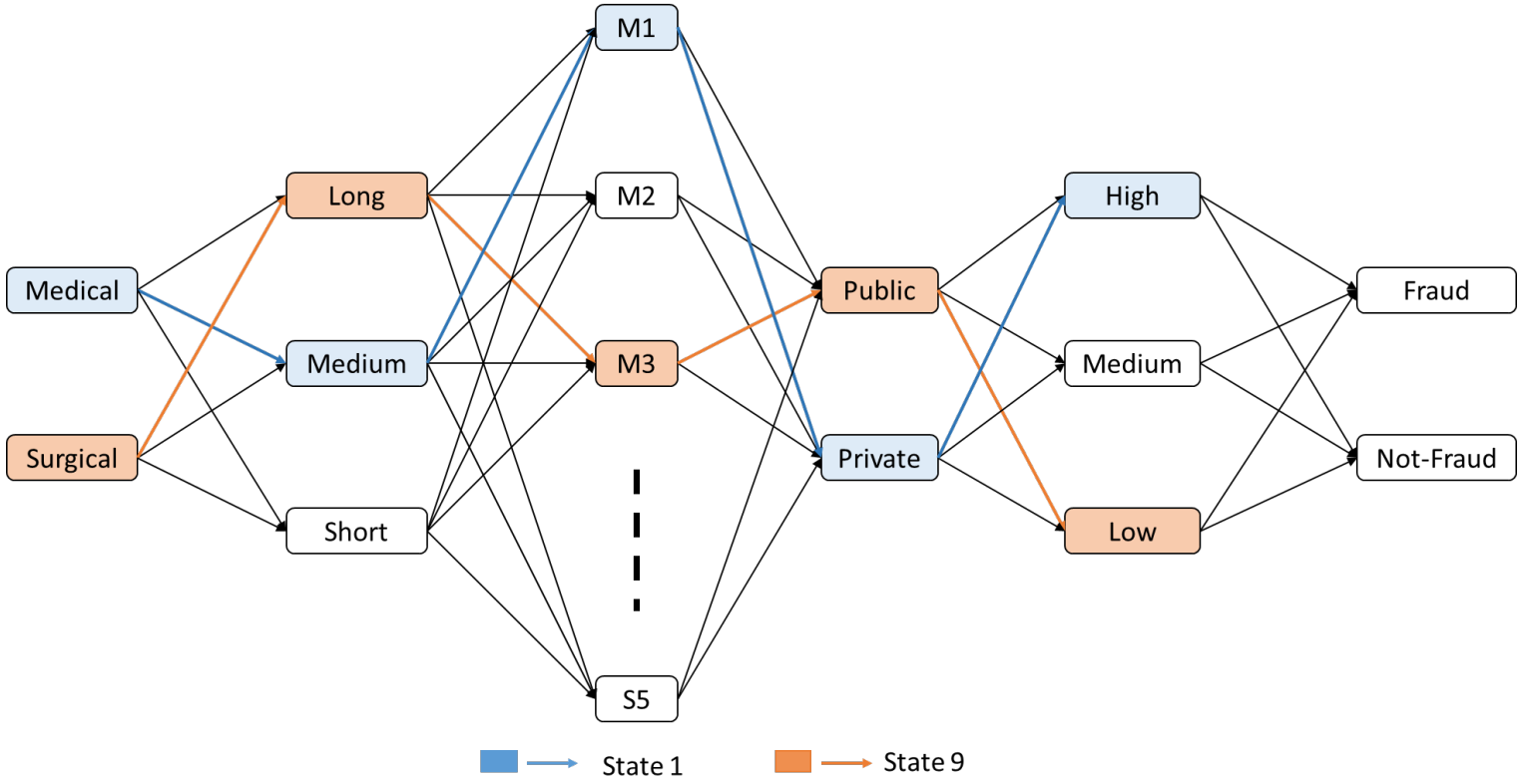
Claims analytics team would further research with new set of additional questions if it is determined to be fraud.

Improved HMM model would monitor further analysis of additional questionnaire and feeds back into the model for better predictions of the future.

Modelling Details

Benefit Type	No of Days Stayed	Primary Diagnosis Code	Hospital Type	Net Amt	States
MEDICAL	medium	M1	Private	high	1
MEDICAL	medium	M1	Private	medium	2
MEDICAL	short	M1	Private	high	3
MEDICAL	medium	M1	Public	medium	4
MEDICAL	long	M1	Public	medium	5
MEDICAL	long	M1	Private	high	6
MEDICAL	medium	M1	Public	medium	4
MEDICAL	long	M3	Private	high	7
MEDICAL	medium	M1	Private	medium	2
MEDICAL	long	M1	Private	high	6
…	…	…	…	…	…
SURGICAL	long	S5	Private	high	8

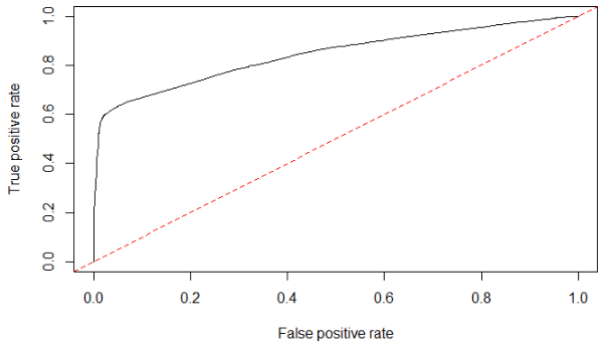
Modelling Details

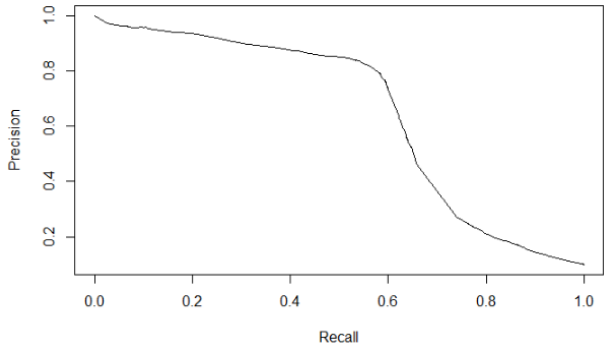


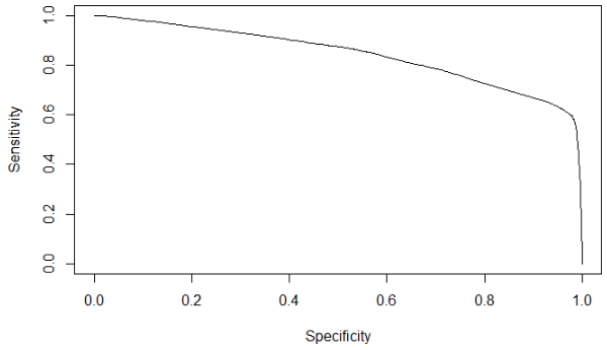
P(Claim = Fraud | State = 1) *= P(Benefit Type = MEDICAL) * P(No of Days Stayed = medium | Benefit Type = medium) * P(Primary Diagnosis Code = M1 | Benefit Type = medium) * P(Hospital Type = Private | Primary Diagnosis Code = M1) * (Net Amt = high | Primary Diagnosis Code = M1) * (Claim Status = Fraud | Primary Diagnosis Code = M1)*

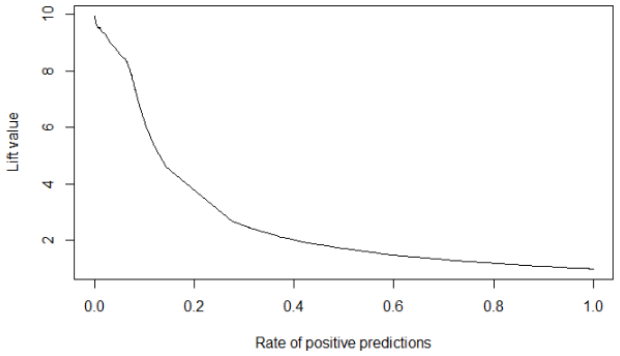
P(Claim = Not-Fraud | State = 1) *= 1 - P(Claim = Fraud | State = 1)*

Results









Measure	Value	Derivations
Sensitivity	0.5940	TPR = TP / (TP + FN)
Specificity	0.9795	SPC = TN / (FP + TN)
Precision	0.7638	PPV = TP / (TP + FP)
Accuracy	0.9407	ACC = (TP + TN) / (P + N)
F1 Score	0.6683	F1 = 2TP / (2TP + FP + FN)

Conclusion

Proposed an HMM application on fraud detection in health insurance domain.

Different steps in claims transaction processing are represented as HMM's underlying stochastic process.

Range of claim amounts are used as observation values and combination of claim characteristics considered as states of HMM

Explained how the HMM can detect whether an incoming claim is fraudulent or not.

THANK YOU

Acknowledgements

- Bhagawan Sri Sathya Sai Baba, Founder chancellor, SSSIHL
- Research team