Institute of Actuaries of India
Statutory body under an Act of Parliament

Expanding the horizon, strengthening the core
20th Global Conference of Actuaries
4th - 6th March, 2019 | Mumbai, India

# Loss Reserving Using ML Techniques

**Parin Kalra**
**Rohit Sahni**

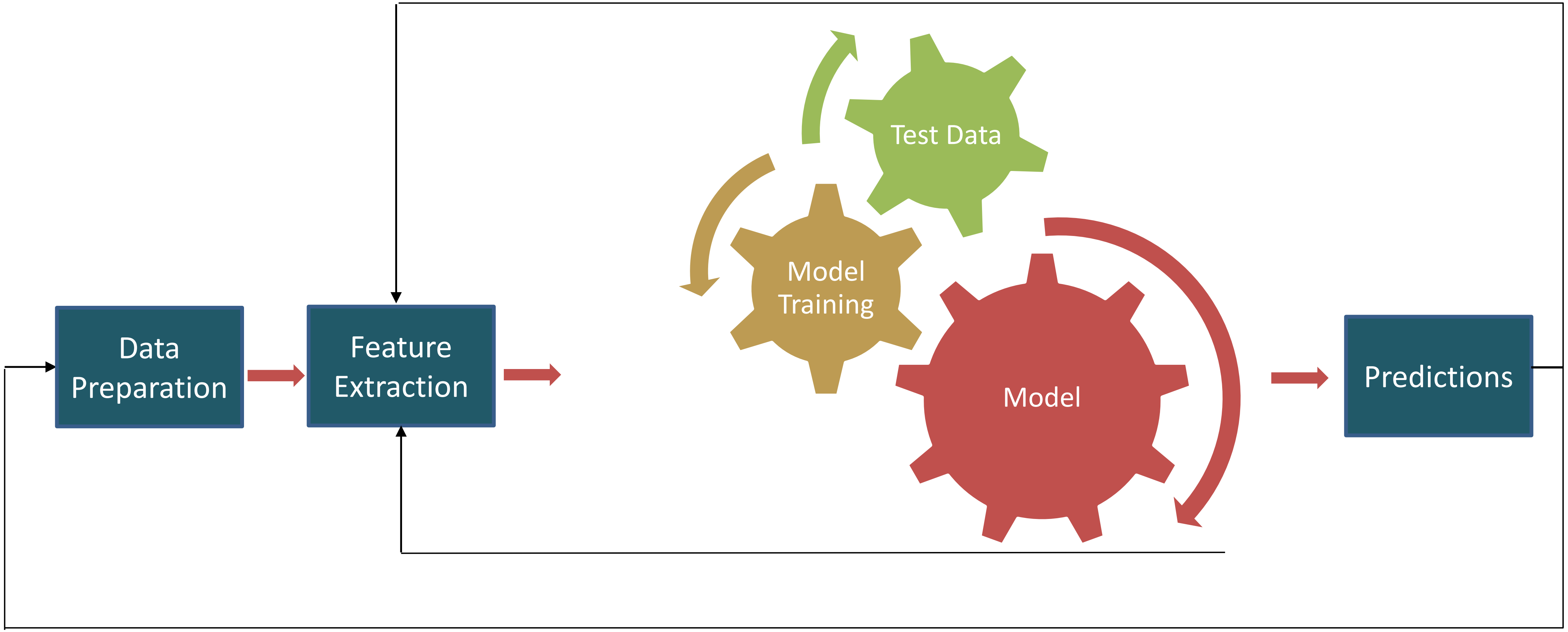
Session –C5
Dated – 05/03/2019

# *Table of Contents*

# Background

- Claims reserving is typically performed on aggregate claim data using familiar reserving techniques such as the chain ladder method & Bornheutter Ferguson Method.
- Rich data about individual claims is often available but is not systematically used to estimate ultimate losses.
- Increasing focus on more insights and better accuracy.
- Do we have techniques to unlock these benefits?

# *Machine Learning – Approach for Next Generation*



# *Approach*

In this presentation we introduce a reserving framework that leverages GBM approach of machine learning to incorporate rich granular information.

## 1 Data Preparation

- Synthetic Data using hypothetical assumptions
- Individual Claim Level Data
- Incorporating fields used in traditional approaches – historical claim payments & incurred amounts
- Claim level characteristics – Driver Age, Driver Gender, Vehicle Value & Vehicle Age

## 2 Modelling

- Applying GBM Technique on training data
- Selecting significant Input variables – Development Lag, Payments to date at previous development lag & Case Estimates to date at previous development lag
- Predicting Ultimate Claim Amount for each Claim

## 3 Validation

- Choosing Validation Data
- Comparing Ultimate predicted with actual ultimate
- Performance measure – Root mean square error
- Approach is repeated on multiple validation data sets

# *Data Structure - Overview*

- Simulated Dataset for **Auto Property Damage claims**
- Sample of 10,000 observations

Data Fields:

| **Claimant Characteristics** | **Dates** | **Amounts** |
|---|---|---|
| Driver's Age, Vehicle Age | Underwriting Date | Payments |
| Driver's Gender | Accident Date | Case Estimate |
| Vehicle Value | Reporting Date | Incurred Amount |

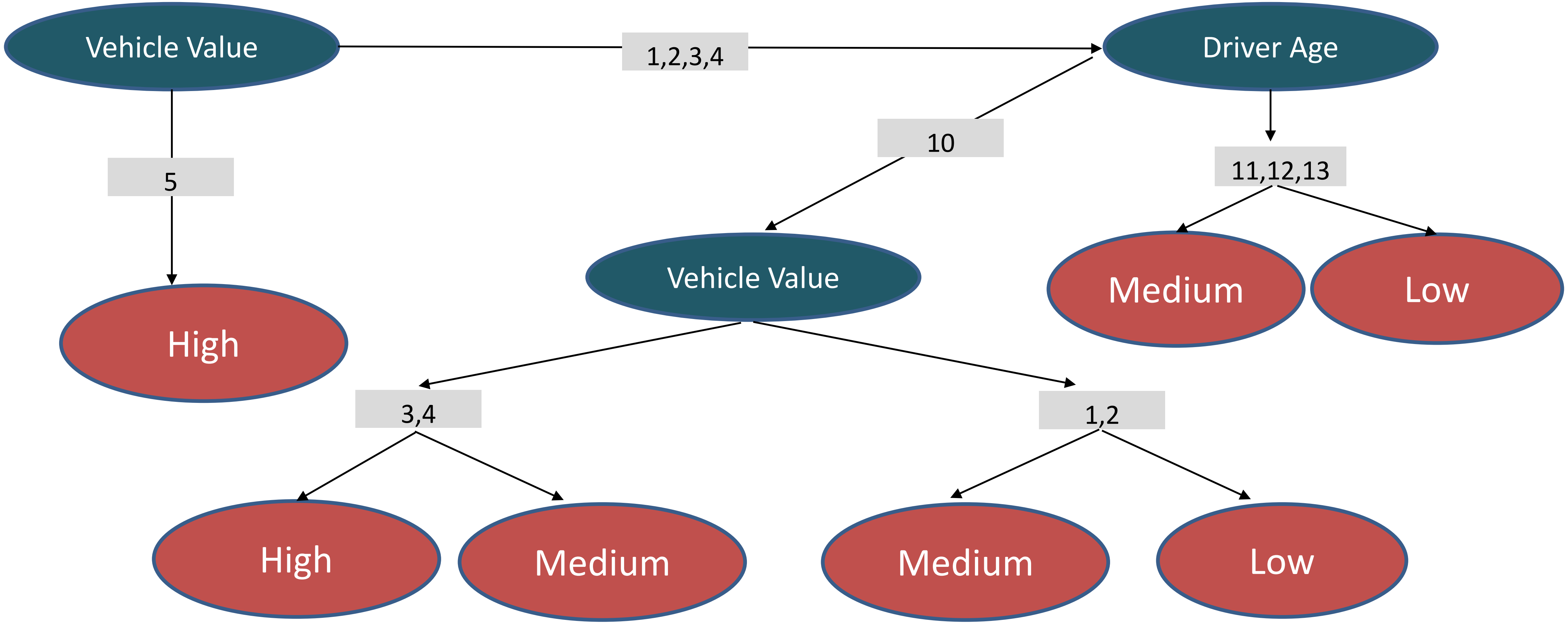
# *Data Structure - Assumptions*

- Claimant Characteristics – Categorical fields distributed randomly

- Dates:
  - Underwriting Date $T_0$ (Randomly Between 2010 & 2015)
  - Accident Date $T_1$ ($T_{0\,+}$ 15 days $<T_1<$ $T_{0\,+}$ 365 days)
  - Reporting Date $T_2$ ($T_1 + 1$ days $<T_2<$ $T_1 + 200$ days, capped till end of 2017)

- Amounts:
  - Incurred Amount at current Lag – Using Decision Tree

# Data Structure - Assumptions




20th Global Conference of Actuaries
4th - 6th March, 2019 | Mumbai, India
Institute of Actuaries of India

# *Data Structure - Assumptions*

- Claim Severity – Log Normal Distribution
- Development Lag – Capped at 5 years

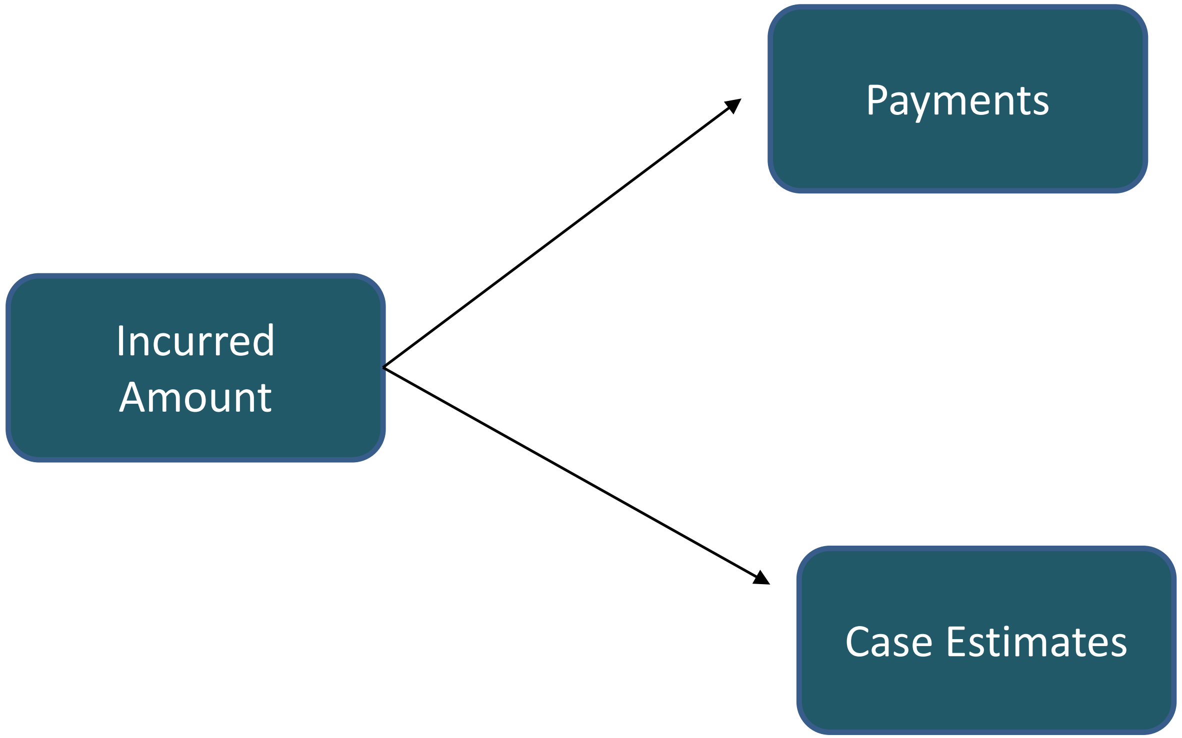
**Industry LDFs**

| Development Lag | Incurred LDFs |
| --- | --- |
| 1 | 1.177036 |
| 2 | 1.133256 |
| 3 | 1.16638 |
| 4 | 0.997724 |
| 5 | 0.997419 |

**Time Series Snapshot**

| Claim Number | Development Lag | Incurred Amount |
| --- | --- | --- |
| 1 | 0 | 10,000 |
| 1 | 1 | 14,000 |
| 1 | 2 | 15,000 |
| 1 | 3 | 14,500 |
| . | . | . |

# *Data Structure - Assumptions*




| Development Lag | Payment Pattern | Payment LDFs |
|---|---|---|
| 0 | 45% - 55% | - |
| 1 | 55% - 65% | 1.522862 |
| 2 | 65% - 75% | 1.189733 |
| 3 | 75% - 85% | 1.176562 |
| 4 | 85% - 95% | 1.001701 |
| 5 | 95% - 100% | 1.001395 |
| >= 6 | 100% | 1.00000 |

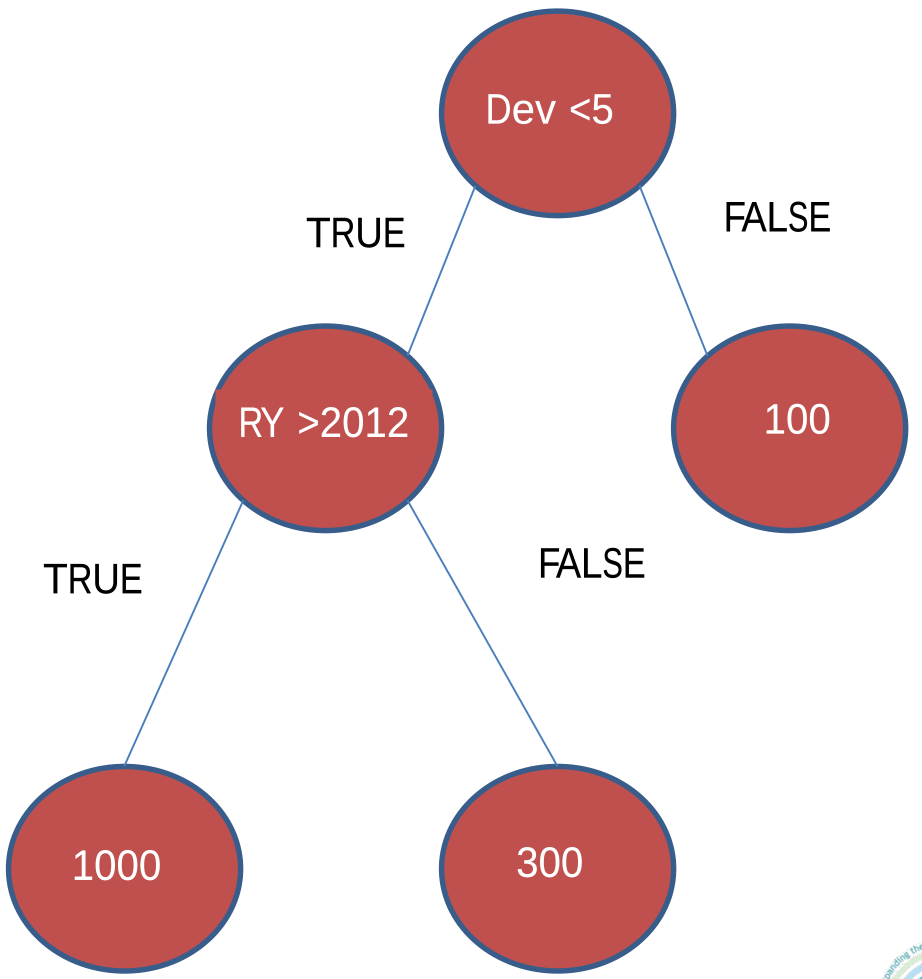
# *Data Structure - Snapshot*

| Claim Number | Driver Age | Driver Gender | Vehicle Value | Vehicle Age | UW Date | Loss Date | Reporting Date | Lag | Payment | Case Estimate | Incurred Amount |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 12 | F | 5 | 6 | 12/9/2014 | 12/9/2014 | 3/4/2016 | 0 | 0 | 63,313 | 63,313 |
| 1 | 12 | F | 5 | 6 | 12/9/2014 | 12/9/2014 | 3/4/2016 | 1 | 35,805 | 39,350 | 75,155 |
| 2 | 10 | M | 4 | 9 | 2/10/2013 | 8/6/2013 | 12/15/2013 | 0 | 0 | 31,959 | 31,959 |
| 2 | 10 | M | 4 | 9 | 2/10/2013 | 8/6/2013 | 12/15/2013 | 1 | 25,030 | 12,907 | 37,937 |
| 2 | 10 | M | 4 | 9 | 2/10/2013 | 8/6/2013 | 12/15/2013 | 2 | 38,117 | 6,535 | 44,653 |
| 2 | 10 | M | 4 | 9 | 2/10/2013 | 8/6/2013 | 12/15/2013 | 3 | 45,350 | 5,254 | 50,603 |
| 2 | 10 | F | 4 | 9 | 2/10/2013 | 8/6/2013 | 12/15/2013 | 4 | 53,357 | 5,666 | 59,022 |
| . | | | . | | | | | | | | |
| . | | | . | | | | | | | | |

# *Modeling- GBM Approach*

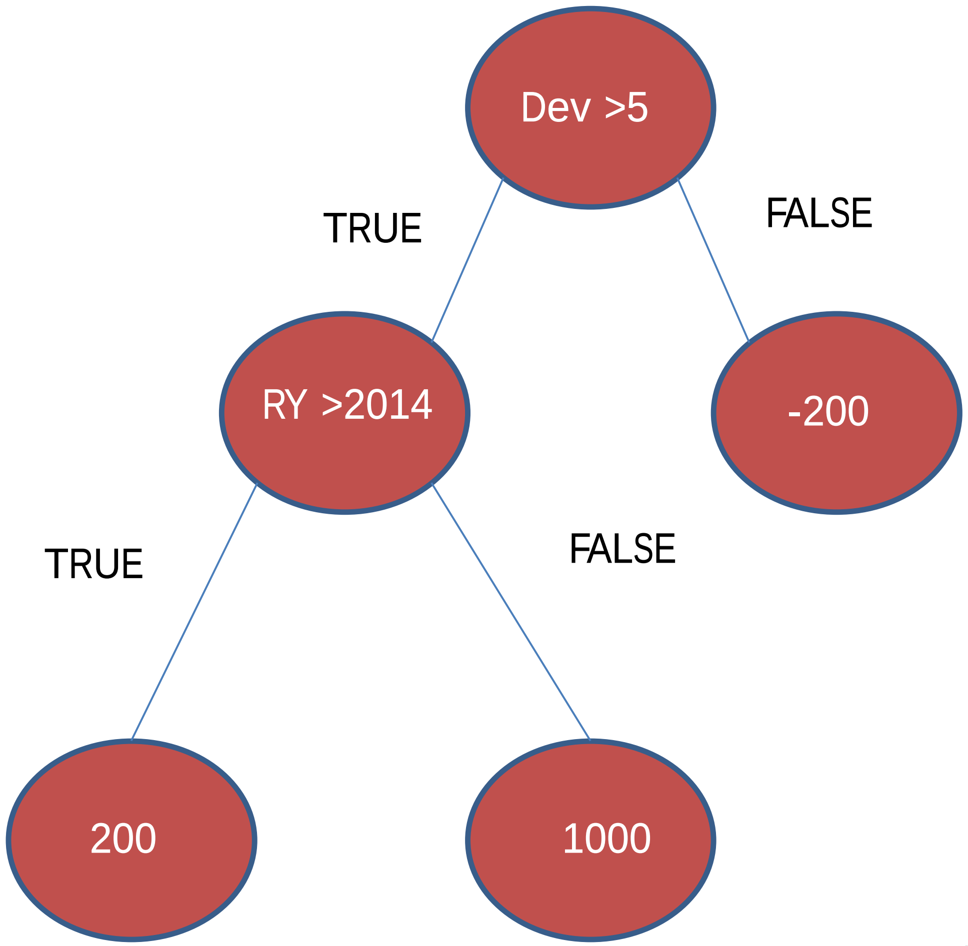
| Report Year | Lag | Actual Values | Predicted Values | Residual |
|---|---|---|---|---|
| 2010 | 7 | 2000 | 900 | 1100 |
| 2012 | 5 | 1562 | 1200 | 362 |
| 2016 | 1 | 30000 | 25000 | 5000 |
| 2010 | 7 | 5600 | 6500 | -900 |
| 2013 | 4 | 2230 | 1800 | 430 |
| 2015 | 2 | 9050 | 8000 | 1050 |
| 2016 | 1 | 6000 | 4500 | 1500 |

Note: The figures used in this slide are purely for representation



# *GBM Approach (Contd.)*

| Report Year | Lag | Residual |
|---|---|---|
| 2010 | 7 | 1100 |
| 2012 | 5 | 362 |
| 2016 | 1 | 5000 |
| 2010 | 7 | -900 |
| 2013 | 4 | 430 |
| 2015 | 2 | 1050 |
| 2016 | 1 | 1500 |

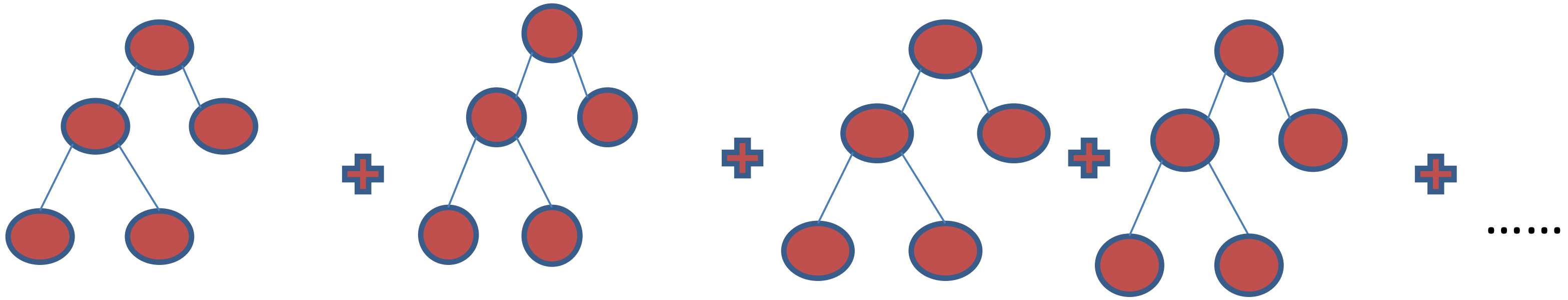



Note: The figures used in this slide are purely for representation

# GBM Approach (Contd.)

| Actual values | Predicted Tree 2 | Residual Tree 1 | Residual Tree 2 | ..... |
|---|---|---|---|---|
| 2000 | 1200 | 1100 | 800 | |
| 1562 | 1262 | 362 | 300 | |
| 30000 | 28000 | 5000 | 2000 | |
| 5600 | 6150 | -900 | -550 | |
| 2230 | 2030 | 430 | 200 | |
| 9050 | 8300 | 1050 | 750 | |
| 6000 | 5100 | 1500 | 900 | |

Note: The figures used in this slide are purely for representation




Each decision tree will reduce the residuals from the previous trees. Therefore, by constructing an efficient number of trees we can minimize the error.

# *Modeling*

| Incurred Amount | Case_lag | Payment_lag | Lag |
|---|---|---|---|
| 59,081 | 293 | 58,941 | 5 |
| 59,234 | 528 | 58,841 | 4 |
| 59,369 | 890 | 50,011 | 3 |
| 50,900 | 2,880 | 42,035 | 2 |
| 44,915 | 10,557 | 27,603 | 1 |
| 38,159 | 32,147 | 0 | 0 |

Dependent variable: Incurred Amount; Independent variables: Case_lag, Payment_lag, Lag

The model predicts incurred at each successive lag leading to an ultimate incurred for each claim.

| Payment | Lag | Predicted Incurred Amount |
|---|---|---|
| 59020 | 5 | 60220 |
| 58900 | 4 | 59180 |
| 51,100 | 3 | 59100 |
| 41790 | 2 | 51000 |
| 27,200 | 1 | 43,400 |

Dependent variable; Independent variables

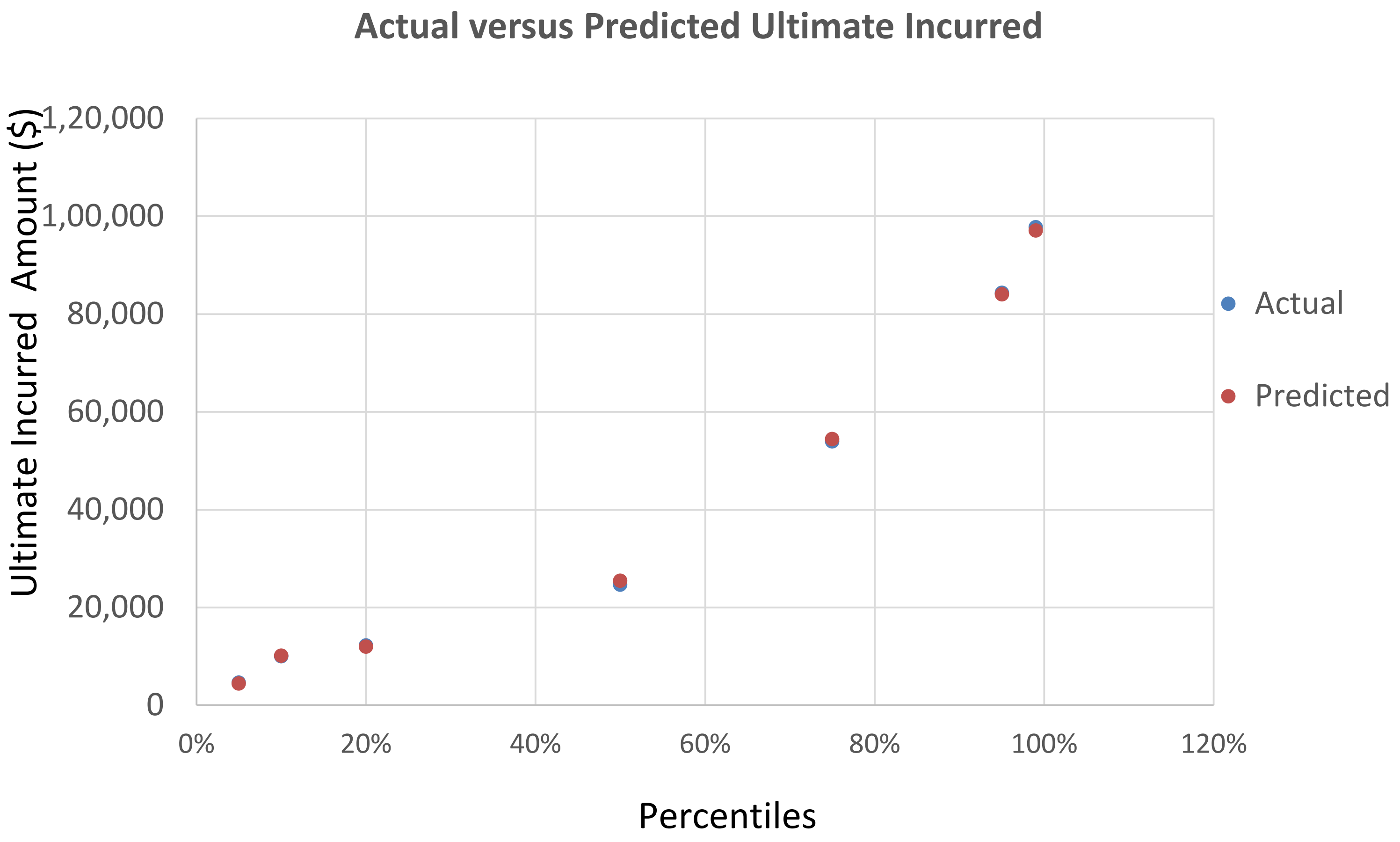
Note: The figures used in this slide are purely for representation

# *Validation & Results*

- Validation Dataset – Multiple data sets (Closed claims, Claims at lag 1)

- Root Mean Square Error (RMSE) of 704 is obtained i.e. on an average predicted deviates from actual ultimate by 704.

- The graph suggests claim size distribution obtained through GBM closely matches the actual distribution (Closed claims).






20th Global Conference of Actuaries
4th - 6th March, 2019 | Mumbai, India
Institute of Actuaries of India

# Validation & Results (contd.)

- Below table compares Ultimates from GBM and Chain Ladder methodologies (Claims at lag 1)

| Predicted | | | |
|---|---|---|---|
| **Report Year** | **GBM** | **Chain Ladder** | **Diff(%)** |
| 2010 | 7,418,500 | 7,410,492 | 0.11% |
| 2011 | 26,038,821 | 26,032,027 | 0.03% |
| 2012 | 28,334,693 | 28,263,852 | 0.25% |
| 2013 | 27,214,391 | 27,204,338 | 0.04% |
| 2014 | 25,108,582 | 25,040,267 | 0.27% |
| 2015 | 23,972,127 | 24,169,841 | -0.82% |
| 2016 | 2,171,166 | 2,232,198 | -2.81% |

## *Way Forward*

- Extend our analysis to new claim types (Bodily Injury, CAT Claims) within auto Line of Business.
- Extend our analysis to other Lines of Business.
- We can consider modeling based on an amalgamation of multiple machine learning techniques as an extension of current GBM model.
- We have not explicitly considered the impact of inflation, claim handling costs, & exposure details in the model.
- How do we account for reopened claims?

Institute of Actuaries of India
Statutory body under an Act of Parliament

Expanding the horizon, strengthening the core

20th Global Conference of Actuaries

4th - 6th March, 2019 | Mumbai, India

# THANK YOU