# Managing Data for Analytics

March 11, 2015

Importance of Predictive Analytics

- Predictive Analytics can help insurers be more effective in all segments of the value chain
  - Marketing – Target and acquire the right customers
  - Actuarial – Prices that accurately reflect risk
  - Underwriting – Select the proper risks and proper products
  - Claims – Identify suspicious claims
- The industry is getting more competitive
  - Top 10 personal auto insurers controlled ½ the market in 1980; now they control 2/3 of the market
  - Only the "fittest" will survive; analytics can provide the needed competitive advantage
  - The industry has recognized the value of analytics

Who Uses Predictive Modeling?

- Predictive analytics is used most often in personal lines.

- 100% of the larger personal lines insurers we surveyed use predictive analytics!

- Of course, personal lines (and PL auto, in particular) is the largest and one of the most competitive segments of the P&C market. Insurers are looking for any competitive edge they can find.

Use of predictive analytics by size of personal lines book

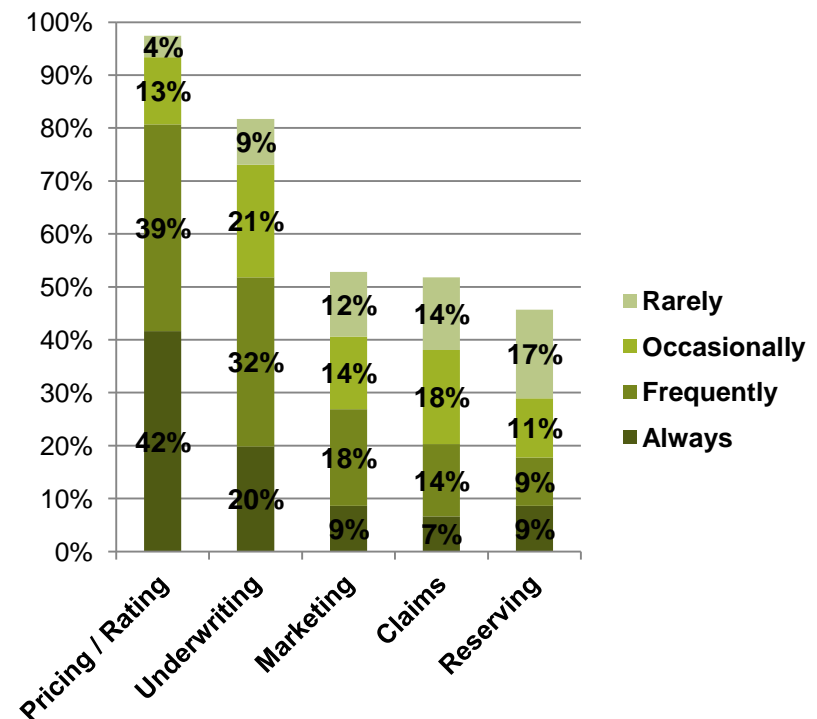< $1B GWP

> $1B GWP

69%

100%
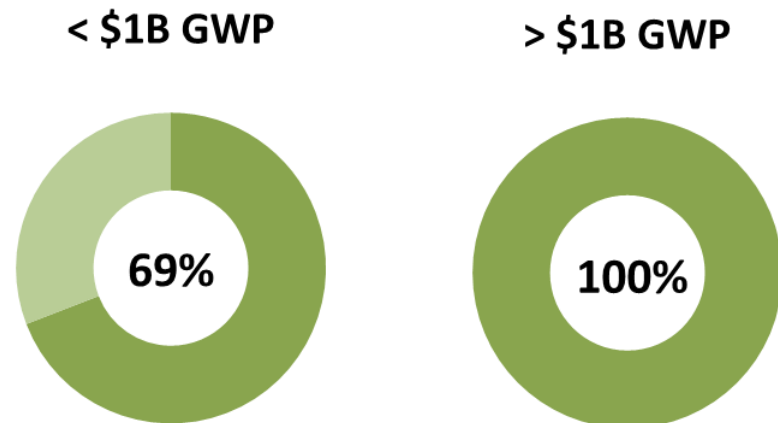
Source: ISO and Earnix 2013 Insurance Predictive Modeling Survey

How Insurers Use Predictive Modeling

- Pricing is the most common use of predictive modeling.

- A majority of insurers also use predictive modeling for underwriting at least frequently.

- But there is still significant usage in marketing, claims, and reserving.

Predictive modeling use by function

Legend:
- Rarely
- Occasionally
- Frequently
- Always

Pricing / Rating: 42%, 39%, 13%, 4%
Underwriting: 20%, 32%, 21%, 9%
Marketing: 9%, 18%, 14%, 12%
Claims: 7%, 14%, 18%, 14%
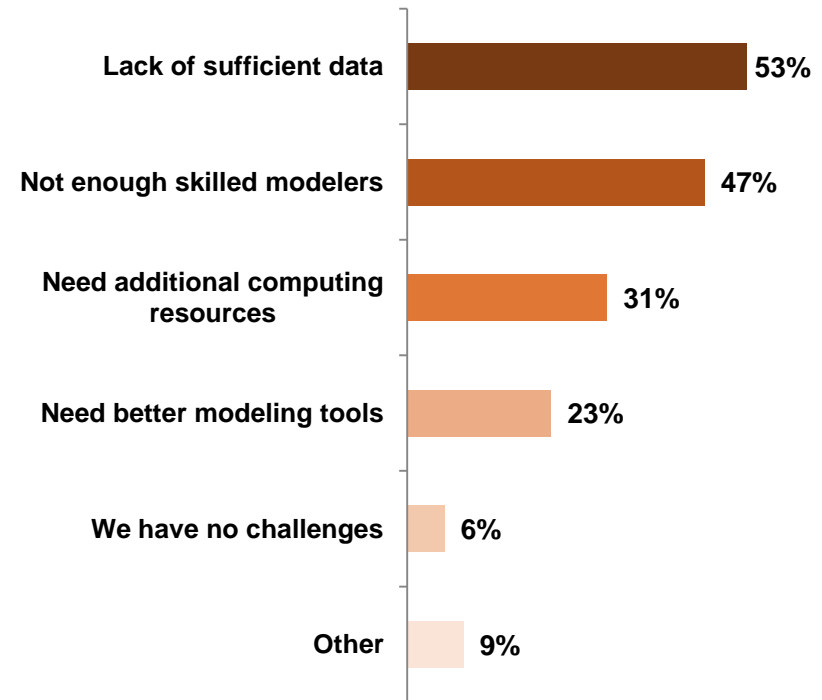Reserving: 9%, 9%, 11%, 17%

Source: ISO and Earnix 2013 Insurance Predictive Modeling Survey

Predictive Modeling Challenges

- Lack of sufficient data is the biggest challenge – both quantity and scope.

- Lack of skilled modelers is a close second most challenging factor for those building an internal predictive modeling capability.

Predictive modeling challenges

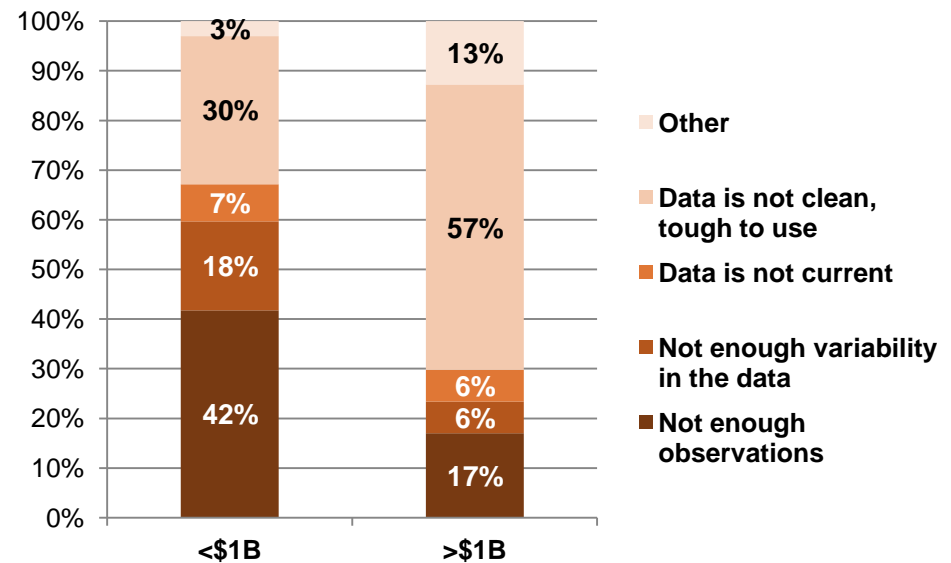| Challenge	Percentage
Lack of sufficient data	53%
Not enough skilled modelers	47%
Need additional computing resources	31%
Need better modeling tools	23%
We have no challenges	6%
Other	9%

Source: ISO and Earnix 2013 Insurance Predictive Modeling Survey

Data Challenges

- Data is a challenge for everybody, but large and small insurers have different challenges.

- Larger insurers are most concerned with data quality.

- Smaller insurers don't have enough observations.

Data challenges by company size

Legend:
- **Other**
- **Data is not clean, tough to use**
- **Data is not current**
- **Not enough variability in the data**
- **Not enough observations**

Chart values:

<\$1B:
- 3%
- 30%
- 7%
- 18%
- 42%

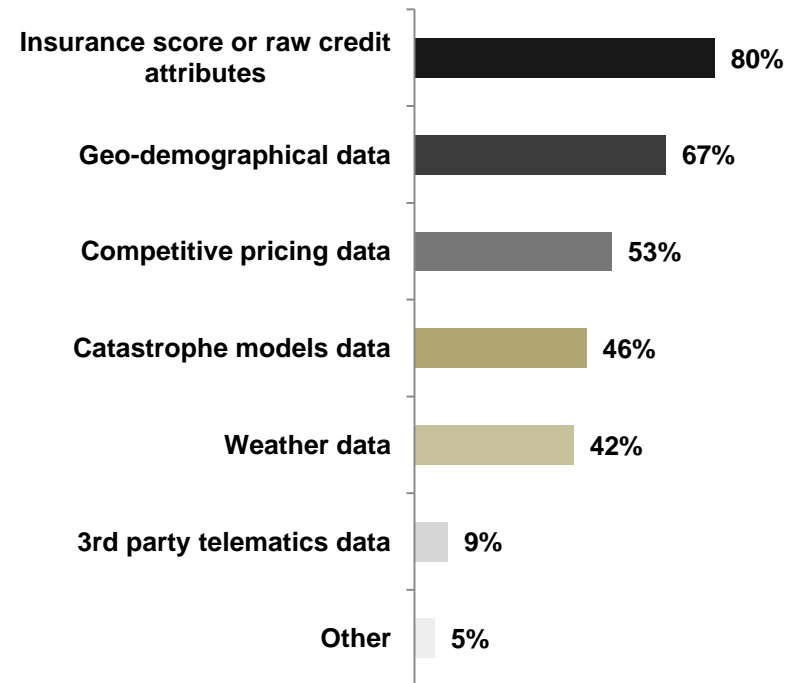>\$1B:
- 13%
- 57%
- 6%
- 6%
- 17%

Numbers may not add up to 100% due to rounding

Source: ISO and Earnix 2013 Insurance Predictive Modeling Survey

Third-Party Data

- More than 90% of insurers supplement their internal data with one or more types of third-party data.

- The most common data types are credit-related data and geo-demographic data.

Types of third-party data used

| Category	Percentage
Insurance score or raw credit attributes	80%
Geo-demographical data	67%
Competitive pricing data	53%
Catastrophe models data	46%
Weather data	42%
3rd party telematics data	9%
Other	5%

Source: ISO and Earnix 2013 Insurance Predictive Modeling Survey

Data Preparation

- Preparing data for analysis is a major bottleneck and drain on resources for most insurers.

- 54% of insurers typically spend more than 3 months to prepare their data for a project.

Data extraction and preparation time
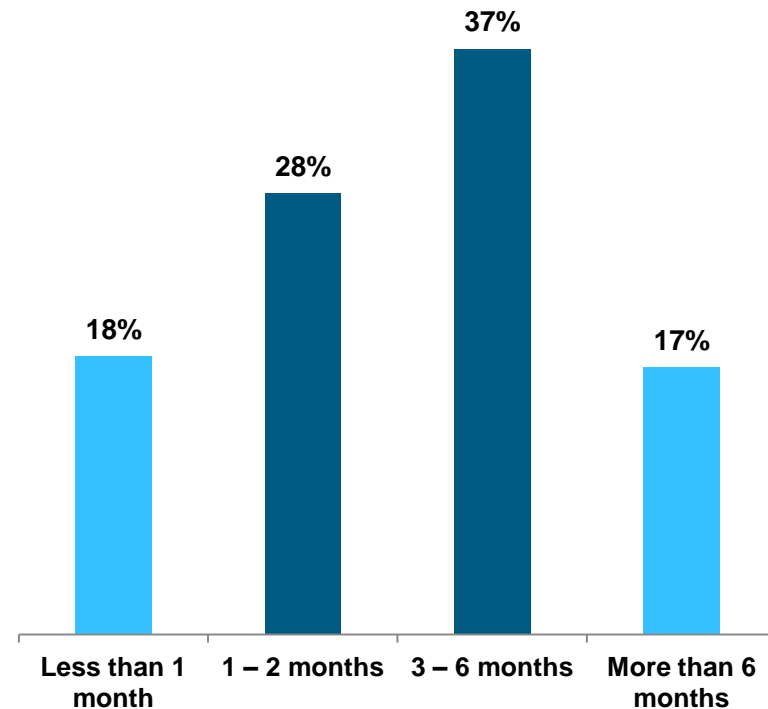
Source: ISO and Earnix 2013 Insurance Predictive Modeling Survey

Good Data: Good Analytics

Good quality data can often compensate for mediocre analysis …

… but, the reverse is never true.

No matter how skilled the analyst, …

… bad data will always lead to bad results!

Data Use for Analytics is Different

- Some Characteristics of Analytics Use of Data
  - Ad Hoc
  - Iterative
  - Sophisticated Users
  - Repurposed Data
  - Denormalized Data

The Data Challenge for Analytics

- Rarely are the operational data stores collected into a single "Enterprise Data Warehouse"
  - You will need to create a useful analytic data store

- Even more rare, is data that has been collected specifically for analytics – usually, analytics is an opportunistic user of data that has been collected for other purposes
  - Data will need to be cleaned, transformed, conformed, and documented before it is certified as "fit for use" for analytics and included in the analytic data store

Insurance Company Data Sources

- Insurance companies collect vast quantities of data in the course of business
- Typical Insurance Analytics Data Sources
  - Customer Relationship Management
  - Quoting/Underwriting
  - Policy Management
  - Billing
  - Claims
  - Audit
  - Actuarial Research
  - Financial Reporting
  - Publically Available Data
  - Third-party data vendors

Considerations for Analytic Database Design

- End goal – Always remember the goal!
  - Two-dimensional flat file for input into modeling software
  - Each record contains an identifier, candidate predictor variables for testing, and one or more target variables
- Analysis requires historical data and the vintage of the predictor variables must be matched to the target variables
- Must support the granularity required for level analysis
- Queries will be ad hoc and iterative

Granularity

- What does each record represent?
- Common record types for insurance analytics
  - Customer-related
    - First named Insured
    - Household
    - Quote
  - Policy
  - Coverage
  - Claim
  - Geography
    - Census tract
    - County
    - State
    - Underwriting Territory
    - Zip Code

Ad Hoc Nature of Data Access

- Few if any queries will be specified at design time – analysts will design their queries as needed
- The iterative nature of the analytic process means the analyst will be back again and again for more and different data

Implications for Analytic Database Design

- Star Schema is often adopted to support analytics
- Data will often be denormalized and aggregated from source systems
- Analytic databases will often grow to contain more history than the source systems. Plan for growth.
- Every variable needs a vintage
- Indexing needs will be imperfectly defined. Count on supporting multiple table joins from any direction = many indexes.
- Granularity – pick the lowest level as your base
  - This means more data, but …
  - … it is the most flexible design. Data can usually be aggregated to a higher level of granularity but you can never go below your base level.

Data Quality and Metadata

- Important Data Quality measures for Analytics:
  - Accuracy
  - Reliability
  - Timeliness
  - Completeness
  - Availability
  - Permissibility
- Analysts usually can't *control* the quality of the data when acquired. So, they must at least *know* the quality of the data in order to determine the usefulness of the data.
- Metadata – documentation of this information

Metadata

- Metadata has many definitions – we mean information about the data that the analyst needs to know in order to use the data appropriately

- Analytic Metadata Needs:
  - Owner/Source
  - Restrictions on use
  - Vintage/update frequency/amount of history
  - Summary statistics
  - Data quality metrics

Conclusion

- Advanced analytics requires different data management support than most other uses of the data

- Two broad areas that demonstrate those needs:

  - Database design

  - Data quality and metadata

- Strive to build an analytic data store that considers the unique needs of analytics.

Questions ?

Phil Hatfield, J.D., CPCU

Head of Modeling Data Services for ISO

phatfield@iso.com