Overview of Big Data, Analytics and Data Science in Insurance

Guy Yollin
Data Science Consultant
Milliman, Inc.

March 1, 2017
Contents

- Big data: from hype to value
- Analytics and predictive modeling
- Data science
- Case study: Variable Annuity Policyholder Behavior
BIG DATA: FROM HYPE TO VALUE
Big data: from hype to value

“Show me the money.”

- Jerry Maguire (1996)
Challenges

- Data that you can’t process and use quickly enough with the technology you have
- Possible reasons for this
  - Volume
  - Velocity
  - Variety (diverse/unstructured formats)
- Not a new problem, but new data sources are increasing the amount of challenging data
Sources of challenging data

- Transactions
- Web log files
- Mobile
- Voice, images, text, video from web and other sources
- Sensors
- Genomic
New data management solutions

- Need to handle larger volumes, unstructured formats, and/or real-time processing have driven new technologies.
- Can lower costs, increase processing speeds for data that can’t be handled well with relational databases and/or single servers.
Opportunities from big data

- Improve models/decisions with
  - new data
  - more data
  - faster cycle times
- Cost reduction
- New products and services
Develop a strategy

- What does your business need?
- What data do you have that is underutilized?
- What data are you missing that would be valuable?
ANALYTICS AND PREDICTIVE MODELING
Business intelligence and analytics

- **Business intelligence** — a set of technologies and tools to understand and analyze business performance
- **Analytics** — the extensive use of data, statistical and quantitative analysis, explanatory and predictive models
- **Predictive analytics** — predicting the value of an outcome, given a number of input measures
Getting more sophisticated = competitive advantage

Optimization
Predictive modeling
Forecasting
Statistical analysis
Alerts
Query/drill down
Ad hoc reports
Standard reports

“What’s the best that can happen?”
“What will happen next?”
“What if these trends continue?”
“Why is this happening?”
“What actions are needed?”
“What exactly is the problem?”
“How many, how often, where?”
“What happened?”

## The methods of analytics

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Machine Learning/Data Mining</th>
<th>Ensembles</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Use assumptions about the probability distribution generating the data to fit model parameters</td>
<td>• Optimize a target function to “learn” complex or unknown patterns in a dataset</td>
<td>• Combine predictions from multiple models to get more accurate predictions</td>
</tr>
<tr>
<td>• Regression and Generalized Linear Models, LARS and LASSO (statistical learning algorithms), survival models, time series</td>
<td>• Decision trees, neural networks, nearest neighbors, naïve Bayes, support vector machines</td>
<td>• Random forests, boosting</td>
</tr>
<tr>
<td>• Principal Components Analysis, Factor Analysis</td>
<td>• Clustering, collaborative filtering, self-organizing map, association rules</td>
<td></td>
</tr>
</tbody>
</table>
Supervised versus unsupervised analytics

**Supervised**

- Predicting the value of an outcome ("target variable"), given other variables
- Examples
  - Regression
  - Decision trees
  - Neural network
  - Nearest neighbors
  - Random forests

**Unsupervised**

- Finding patterns in data, but not predicting a specific target variable
- Examples
  - Clustering
  - Anomaly detection
  - Association rules
  - Hidden Markov
  - Self-organizing map
# Predictive modeling techniques compared

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Machine Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis testing</td>
<td>Hypothesis generation</td>
</tr>
<tr>
<td>Small datasets</td>
<td>Big datasets</td>
</tr>
<tr>
<td>Uncertainty</td>
<td>Predictive accuracy</td>
</tr>
<tr>
<td>Distributions</td>
<td>Rules and algorithms</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>CART, Random Forest, Boosting</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>SVM</td>
</tr>
<tr>
<td>Generalized Linear Models</td>
<td>Neural Nets</td>
</tr>
<tr>
<td>Generalized Additive Models</td>
<td>K-nearest neighbors</td>
</tr>
</tbody>
</table>
Model selection and overfitting

- It is important to assess models on an independent (“holdout”) sample, or at least use methods for model selection that “penalize” for complexity.

- While it might appear that a complex model fits the data well, if it is fitting “noise” in the training sample it will not work as well when applied on new observations.
Comparing popular methods for predictive/supervised models

Linear methods
- Function to relate input variables to target variable
- Greater differentiation
- Harder to interpret in aggregate, but can provide “reasons” for individual predictions

Decision trees
- Partition data into mutually exclusive groups
- Easy to understand, multivariate insights
- Cannot split too much before you run out of data

Ensembles
- Fit many “learners” on different random samples of the data and combine their predictions
- Greatest differentiation
- Hardest to interpret reasons for predictions
## Infrastructure for advanced analytics

### Analysis software

<table>
<thead>
<tr>
<th>Analysis software</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS, SPSS</td>
<td>Strong at data manipulation</td>
<td>Cost</td>
</tr>
<tr>
<td></td>
<td>Widely used by large enterprises</td>
<td>Steep learning curve</td>
</tr>
<tr>
<td>R</td>
<td>Open source, low cost</td>
<td>Limited but growing enterprise support</td>
</tr>
<tr>
<td></td>
<td>Cutting-edge techniques available sooner</td>
<td></td>
</tr>
<tr>
<td>In-database/MapReduce</td>
<td>Fast – do not move data out of database, and use processing power on the database servers</td>
<td>Limited set of algorithms</td>
</tr>
<tr>
<td>Scripting languages (i.e., Python)</td>
<td>Open source, low cost Strong data manipulation with numpy/pandas</td>
<td>Steep learning curve</td>
</tr>
</tbody>
</table>

### Data management

<table>
<thead>
<tr>
<th>Data management</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relational database (e.g. Oracle)</td>
<td>Widely used Good for structured data</td>
<td>Doesn’t scale well for big volume Poor with unstructured data</td>
</tr>
<tr>
<td>Massively parallel (e.g. Oracle Exadata, Teradata)</td>
<td>Extends volume Good for structured data</td>
<td>More expensive than NoSQL solutions Poor with unstructured data</td>
</tr>
<tr>
<td>NoSQL/Non-relational (e.g. Hadoop)</td>
<td>Scales across cluster Handles unstructured data well</td>
<td>Learning curve</td>
</tr>
</tbody>
</table>
Evolving predictive models

Often new data produce larger gains than changing to more sophisticated mathematics.
Gartner Hype Cycle for Emerging Technologies
Gartner Hype Cycle for Emerging Technologies
Skills for predictive modeling
Cross-industry standard data mining process (CRISP-DM)
Summary

- Predictive analytics use data to determine the probability of a future outcome
- Predictive analytics can be used to improve a wide variety of decisions and processes
- Predictive analytics can be improved both through better data and/or better modeling techniques
- Predictive analytics require skills in data manipulation, programming, statistics/machine learning, and project management
DATA SCIENCE IS THE ANSWER, WHAT’S THE QUESTION
Some definitions of data scientist

- A data analyst in California
- A statistician under 35
- A developer of “data products”
- A practitioner of “data jujitsu”
- Knows more statistics than a software engineer and knows more software engineering than a statistician
Something new, or re-branding?

C. F. Jeff Wu (1998):
- Data collection
- Modeling and analysis
- Problem solving and decision making

William S. Cleveland (2001):
- Multidisciplinary investigation
- Models and methods
- Computing with data
- Tool evaluation

Great article: 50 Years of Data Science by David Donoho
http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf
Some more recent attempts

The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it

Combine the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data

start by looking at what the data can tell them, and then picking interesting threads to follow, rather than the traditional scientist’s approach of choosing the problem first and then finding data to shed light on it

Extract information from large datasets and then present something of use to non-data experts
What seems different

- Using large datasets
- Hands-on, heavy data prep of unstructured data
- Coding with general purpose languages (Python, C++, Java)
- Starting with the data, not a question?
- Emphasis on storytelling/visualization
The data science Venn diagram

- Data wrangling
- Programming and software development
- Statistics, machine learning, and data mining
- Data visualization and story-telling

Domain knowledge
- Ability to ask good questions and break down problems
- Ability to define and test hypotheses
MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand what a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS
★ Machine learning
★ Statistical modeling
★ Experiment design
★ Bayesian inference
★ Supervised learning: decision trees, random forests, logistic regression
★ Unsupervised learning: clustering, dimensionality reduction
★ Optimization: gradient descent and variants

PROGRAMMING & DATABASE
★ Computer science fundamentals
★ Scripting language e.g. Python
★ Statistical computing package e.g. R
★ Databases SQL and NoSQL
★ Relational algebra
★ Parallel databases and parallel query processing
★ MapReduce concepts
★ Hadoop and Hive/Pig
★ Custom reducers
★ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS
★ Passionate about the business
★ Curious about data
★ Influence without authority
★ Hacker mindset
★ Problem solver
★ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION
★ Able to engage with senior management
★ Story telling skills
★ Translate data-driven insights into decisions and actions
★ Visual art design
★ R packages like ggplot or lattice
★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau
Managing big data

“You’re gonna need a bigger boat.”

- Jaws (1975)
Managing big data

- Distribute data storage, data processing across multiple computers
- Can use cheaper, commodity hardware because data is duplicated on multiple machines – can be recovered when one fails
- Faster run times - use the parallel computing power of the machines where the data is stored, and avoid I/O of extracting data first
Let’s talk about the elephant in the room, Hadoop

- Software framework for storing and processing structured and unstructured data
- Distributes (and replicates) your data across multiple commodity machines (a “cluster”)
- File system (HDFS) keeps track of where the data is
- Programming framework (MapReduce) to process the data
The cloud makes Hadoop much more accessible

### Hadoop in the Cloud - Options

#### Hadoop in IaaS
- **Pros**
  - Complete Control
  - On-Demand Cluster Sizing
  - Storage - Local or Cloud
- **Cons**
  - Only VMs managed for HA
  - Administration required
  - Clusters need to stay active

#### Hadoop in PaaS
- **Pros**
  - Fully managed – SLA bound
  - Flexible resizing
  - Pay-on-use
  - Customization Options
  - Deployed in minutes
- **Cons**
  - Forgo some control

#### Big Data as a Service
- **Pros**
  - Abstracted from clusters
  - Automated resource alignment
  - Easy to use interface and APIs
  - Familiar languages
- **Cons**
  - Forgo complete control
  - Priced extravagantly
## “Not Only Hadoop”

<table>
<thead>
<tr>
<th>Family</th>
<th>Category</th>
<th>Examples</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relational</td>
<td>Massively Parallel Processing (MPP)</td>
<td>Teradata, Netezza, Greenplum, Vertica, Oracle Exadata</td>
<td>Fast and familiar</td>
<td>Expensive Poor for unstructured data</td>
</tr>
<tr>
<td>“Not Only SQL”</td>
<td>Key-Value</td>
<td>Redis, Riak, Voldemort</td>
<td>Simple, fast I/O</td>
<td>Poor for complex data</td>
</tr>
<tr>
<td></td>
<td>Column</td>
<td>Hbase, Hypertable, Cassandra</td>
<td>Good for unstructured data</td>
<td>Poor for interconnected data</td>
</tr>
<tr>
<td></td>
<td>Document</td>
<td>CouchDB, MongoDB</td>
<td>Good for unstructured data</td>
<td>Poor for interconnected data</td>
</tr>
<tr>
<td></td>
<td>Graph</td>
<td>Neo4j, InfiniteGraph</td>
<td>Certain types of problems</td>
<td>Not really scalable</td>
</tr>
</tbody>
</table>
Analyzing big data

“I feel the need – the need for speed!”

- Top Gun (1986)
First, it isn’t always as big as it seems

- Use big data tools to summarize it down, then apply the usual analysis software
- Do you really need every observation? Then sample it down
Intermediate steps

- Use software/algorithms that process outside of memory (bigGLM, Revolution R)
- Get more memory – a new machine, a big memory instance on a cloud
If you go for it . . .

Need analysis software that has been written to work in parallel

<table>
<thead>
<tr>
<th>Product</th>
<th>Algorithms supported for distributed processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS on Hadoop</td>
<td>C&amp;RT, Time series, GLM, Logistic regression, Random Forest, Clustering</td>
</tr>
<tr>
<td>Revolution R Enterprise</td>
<td>Regression, Logistic regression, GLM, Clustering, Decision Trees, Random Forest</td>
</tr>
<tr>
<td>IBM SPSS Analytic Server</td>
<td>Linear regression, Neural Net, C&amp;RT, CHAID</td>
</tr>
<tr>
<td>Mahout</td>
<td>Collaborative filtering, Naïve Bayes, Random Forest, Clustering, Principal Components</td>
</tr>
<tr>
<td>MapReduce</td>
<td>Write your own MapReduce directly or with an interface like RHadoop</td>
</tr>
</tbody>
</table>
APPLICATIONS TO INSURANCE
Wherever decisions are made, there is opportunity for predictive modeling

- **Marketing**
  - Brand management
  - Target marketing
  - Cross sell
  - Product design

- **Underwriting**
  - Underwriting requirements
  - Exposure audits

- **Pricing**
  - Rate relativities

- **Claims**
  - Fast track
  - High risk case management
  - Fraud

- **Distribution**
  - Agency selection
  - Agency management
Wherever predictive modeling is used, there is an opportunity for new data.
Example: Property & Casualty Pricing

Univariate
- Fewer variables
- Loss ratio relativities

GLMs
- More variables, interactions
- Frequency, severity, pure premium

Machine Learning
- Greater predictive power for larger datasets
- Less interpretability

Optimization
- Solving for rating factors given expected loss cost and elasticity models

1990s

2000s

2010s
Life/annuity business functions that can benefit from predictive modeling

<table>
<thead>
<tr>
<th>Customer Acquisition</th>
<th>Pricing</th>
<th>Financial Risk Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Propensity to convert</td>
<td>• Product development</td>
<td>• Reserving and capital management</td>
</tr>
<tr>
<td>• Expected profitability/retention</td>
<td>• New business pricing</td>
<td>• Hedging</td>
</tr>
<tr>
<td>• Triaged underwriting</td>
<td>• Renewal rate setting</td>
<td>• Financial reporting</td>
</tr>
<tr>
<td>• Website customization</td>
<td>• Experience monitoring</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Operations</th>
<th>Sales and distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Policy servicing</td>
<td>• Recruiting</td>
</tr>
<tr>
<td>• Claims processing</td>
<td>• Selection</td>
</tr>
<tr>
<td>• Fraud detection</td>
<td>• Management</td>
</tr>
</tbody>
</table>
CASE STUDY:
VARIABLE ANNUITY
POLICYHOLDER BEHAVIOR
Milliman Life & Annuity Predictive Analytics Team

Joint team with diverse backgrounds:
- Data scientists and statisticians
- Data managers
- Actuaries and subject-matter experts
- Business strategists
- Technology specialists
Variable annuity lapse behavior

- **Market Size:** There is nearly $2 Trillion in AUM in the VA market.

- **Guarantees:** A significant portion of VA’s sold in the last decade has long-term guarantees (Guaranteed Living Withdrawal Benefits).

- **Uncertainty:** There is enormous uncertainty around how policyholder behavior, particularly lapse behavior, will emerge over the lifespan of business currently on the books of companies.

- **Impact:** The impact of policyholder behavior on the value and profitability of VA business is enormous, both for the industry as a whole and for individual companies. The impact could be in the billions of dollars for the companies with the largest exposure, and potentially long-term solvency. **Incremental improvements to lapse modeling can have large dollar impacts.**

- **Availability of data:** For companies who have been consistently present in the VA marketplace, there is now over a decade of experience. **While the experience data has some meaningful limitations in forecasting future experience, the industry could likely gain significant value by using the available data to develop better forecasting tools.**
Current Policyholder Behavior Paradigm

- Market risks can be hedged, but policyholder behavior is more difficult to estimate and therefore difficult to hedge
- Higher than expected persistency during the financial crisis caused some large insurers to take significant charges

“Moody’s expects that if interest rates remain low, equities markets fall, and guarantees stay in-the-money, similar behavior by policyholders will continue. This will force companies to take charges in recognition of lower prospective profitability... the industry impact could be in the billions of dollars...”

-Moody’s, June 2013

“... the hedging of variable annuity risk requires the company to make policyholder behavior assumptions that may prove inaccurate. Deviations from pricing and hedging assumptions could have a material negative impact on [a major insurer’s] capital and earnings in a severe, albeit unexpected, scenario.”

-Fitch, December 2014
Goals of the Study

- Establish more accurate expectations of VA lapses
  - Discover new predictors of lapse behavior
  - Investigate macroeconomic effects
  - Fill in “data holes”
- Create an industry lapse benchmark by rider type
- Enable comparison to industry experience
- Enable discussion of current industry lapse models
Potential Predictor Variables

300+ variables cover many types of data

- Policyholder attributes
- Past policyholder behavior
- Policy and rider features
- Policy values
- Macroeconomic factors
Predictive Modeling Process

Data Preparation

VALUES

1300 files

117 million records

300+ variables

Data Analysis

Model Building

1300 files

VALUES

117 million records

300+ variables

training/holdout test
Lapse
The most comprehensive and rigorous examination of VA industry lapse experience
**Lapse**

New drivers behind lapse

- Previous behavior matters – e.g., withdrawal behavior is a major driver of lapse
- Demographics matter (age, gender)
- Product design matters
  - Different predictors for different guarantees
  - Different sensitivities for different richness
- Macroeconomic conditions help explain behavior
Lapse
Adding variables improve prediction

GMAB
GWB
GMIB
No LB

Models
- Baseline PM with Forward-looking ITM
- VALUES Predictive Model

Actual to Expected Ratio vs. Rank of relative probabilities
Use predictive analytics to answer specific questions

Q: Does sensitivity to the spread vary by distribution channel?

A: Yes. And we can quantify them.
Use predictive analytics to discover non-linear relationship

FA: Cash surrender value

Relation of lapses to cash surrender value is fit best by a combination of linear and quadratic components.
Thought Experiment

Is it possible that we would better be able to predict the behavior of individual policyholders if we knew about……

- … their life events (death of spouse, marriage, divorce, events in lives of their children)
- … how their financial assets and investments are performing?
- … whether they are current on their mortgage or have paid it off?
- … whether they are in good health or have health problems?
- … if they have made big recent purchases?

Much of this information may be readily available.

There might be other available data that serve as useful proxies for this information.
Vision

Traditional State

Where Industry is

Predictive Modeling State

Big Data State

- Traditional one-way actuarial techniques to estimate behavior by age/duration and limited number of other characteristics using experience where it exists
- Primarily macro-oriented... little use of detailed information on policyholder characteristics
- Judgment and guesswork where experience does not exist

- Next-generation experience studies using policyholder longitudinal data.
- Use much wider set of explanatory variables readily available to company
  - Internal data (Product features, distribution channel, policyholder and contract characteristics)
  - Macro data (Economic data, financial market conditions)
- More sophisticated analysis techniques to find non-linear, multivariate effects, complex interactions

- Employ external consumer/financial/health and big/unstructured data sources in a full Predictive Analytics framework.
- Develop individual policyholder profiles
Data enrichment: Sources of external data

- Credit – credit scores, attributes of credit data
- Lifestyle – shopping, interests, recreation
- Health – relative mortality score, e.g.
- Mortgage – history of current mortgage, property attributes
- Census – demographic data by census block

* Icons made by Freepik and Madebyoliver from www.flaticon.com
## Value of progressing to advanced state

### New Capabilities

<table>
<thead>
<tr>
<th>Improved Estimates of Future Lapse Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification of Policyholder “Value Profiles”</td>
</tr>
<tr>
<td>Real-time Data on Policyholder Activity</td>
</tr>
<tr>
<td>Reusable Framework for Other Customer Behavior Risks</td>
</tr>
</tbody>
</table>

### Potential Actions/Benefits

- **Hedging** better targeted to true liabilities, with improved tracking
- **Reserving** better aligned with actual liabilities, possible release
- **Capital Management** -- potentially more optimal use of capital
- Product development tailored to customer segments
- Pricing new products based on better information on customer behavior
- Inducements and offers to existing customers based on profiles
- Continuous monitoring of policyholders using up-to-date data and refreshed models
- Identification of “at-risk” policyholders for potential action
- Improved modeling and monitoring of guarantee exercise/utilization
- Improved modeling and monitoring of inefficiency of withdrawals
The data product Venn diagram

Innovation

- Proprietary web-based interactive platform for our clients to view data and results
- Designed for industry study participants
- Adaptable for company’s own purpose
On modeling

“All models are wrong but some models are useful.”

- George E. P. Box (1978)
Questions?

Guy.Yollin@milliman.com