CENTER for
ECONOMIC
JUSTICE

FAIR ACCESS
FAIR TREATMENT

Predictive Analytics in Insurance: Regulatory Oversight Needed

Presentation to
NAIC Big Data Working Group

Birny Birnbaum
Center for Economic Justice

August 26, 2016

The Center for Economic Justice

CEJ is a non-profit consumer advocacy organization dedicated to representing the interests of low-income and minority consumers as a class on economic justice issues. Most of our work is before administrative agencies on insurance, financial services and utility issues.

On the Web: www.cej-online.org

Why CEJ Works on Insurance Issues

Essential Financial Security Tool for Individual and Community Economic Development: CEJ Works to Ensure Access and Fair Prices for These Essential Products and Services, particularly for Low- and Moderate-Income Consumers.

Primary Institution to Promote Loss Prevention and Mitigation: CEJ Works to Ensure Insurance Institutions Maximize Their Role in Efforts to Reduce Loss of Life and Property from Catastrophic Events.

Big Data Defined

Insurers' use of Big Data has transformed the way they do marketing, pricing and claims settlement. Big Data means:

- Massive databases of information about (millions) of individual consumers
- Associated data mining and predictive analytics applied to those data
- Scoring models produced from these analytics.

The scoring models generated by data mining and predictive analytics are algorithms. Algorithms are lines of computer code that rapidly execute decisions based on rules set by programmers or, in the case of machine learning, generated from statistical correlations in massive datasets. With machine learning, the models change automatically.

Barocas and Selbst: *Big Data's Disparate Impact*

"In contrast to those traditional forms of data analysis that simply return records or summary statistics in response to a specific query, data mining attempts to locate statistical relationships in a dataset. In particular, it automates the process of discovering useful patterns, revealing regularities upon which subsequent decision-making can rely. The accumulated set of discovered relationships is commonly called a "model," and these models can be employed to automate the process of classifying entities or activities of interest, estimating the value of unobserved variables, or predicting future outcomes.

Barocas and Selbst: *Big Data's Disparate Impact (con't)*

"These all involve attempts to determine the status or likely outcome of cases under consideration based solely on access to *correlated* data. Data mining helps identify cases of spam and fraud and anticipate default and poor health by treating these states and outcomes as a function of some other set of observed characteristics.

In particular, by exposing so-called "machine learning" algorithms to examples of the cases of interest (previously identified instances of fraud, spam, default, and poor health), the algorithm "learns" which related attributes or activities can serve as potential proxies for those qualities or outcomes of interest. In the machine learning and data mining literature, these states or outcomes of interest are known as "target variables."

Barocas and Selbst: *Big Data's Disparate Impact (con't)*

"The proper specification of the target variable is frequently not obvious, and it is the data miner's task to define it. In doing so, data miners must translate some amorphous problem into a question that can be expressed in more formal terms that computers can parse.

Through this necessarily subjective process of translation, though, data miners may unintentionally parse the problem and define the target variable in such a way that protected classes happen to be subject to systematically less favorable determinations."

Examples of Insurer Big Data Algorithms

Pricing:

- Price Optimization/Demand Models
- Customer Value Scores,
- Telematics,
- Credit Scores,
- Criminal History Scores,
- Vehicle Scores,
- FireLine Rating

Claims:

- Fraud Scores,
- Severity Scores

Personal Consumer Information in Big Data

- Social Media
- Shopping Habits/Purchase History
- Hobbies and Interests
- Demographics/Household Data/Census Data
- Government Records/Property Records
- Web Tracking
- Mainstream Credit Files: Loans, Credit Cards
- Alternative Credit Data: Telecom, Utility, Rent Payment
- Telematics / Wearable Devices

Consumer Protection / Regulatory Oversight Needed

1. Correlation is Not Causation / Spurious Correlation / Post-Hoc Hypothesis Testing
2. Predictive Models May Reflect and Perpetuate Historic Discrimination

   a. Biased Data

   b. Biased Model / Assumptions

   c. Faulty Model Specification --

Spurious Correlation

A spurious correlation is a statistically-valid association between variables that is not causally related.

Data Mining and Big Data Models – particularly in insurance – are premised on correlation, not causation.

See http://www.tylervigen.com/spurious-correlations

U.S. Spending on Science & Suicides by Hanging: 99.7%

Maine Divorce Rate & Per Capita Margarine Consumption: 99.3%

Per Capita Mozzarella Consumption & Engineering PhDs: 95.9%

Post-Hoc Hypothesis Testing – justifying the results of data mining. Example: Insurance credit scoring works because . . ..

Warning by the American Statistical Association

"While the p-value can be a useful statistical measure, ***it is commonly misused and misinterpreted***" (page 7).

ASA Statement on Statistical Significance & P-Values. Feb. 5, 2016. Ronald L. Wasserstein, Executive Director, ed., on behalf of the American Statistical Association Board of Directors.

Statement by the American Statistical Association

Proper inference requires full reporting and transparency.

“Cherry-picking promising findings, also known by such terms as data dredging, significance chasing…and ‘p-hacking,’ leads to a spurious excess of statistically significant results…and should be vigorously avoided.”

“Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted and all p-values computed.”

Example: Claim Fraud Scores, Claim Severity Scores

LexisNexis Claim Tools

“LN has over 10,000 data sources that feed into its infrastructure each month and has contributed information from the industry.

“Claims Data Fill” – deliver data and analytics directly into claims system in the claims process regarding parties, vehicles and carrier information. Used to verify information provided to insurers and provide indicators beyond the data to identify whether a social security number is an indicator of fraud or whether an address provided is a good address. Has an analytic component at first notice of loss and throughout the claim, constantly monitoring the claim looking for fraudulent activities. Real time data verification and enhancement with fraud scoring and attributes

LexisNexis Claim Tools (con’t)

“Example, insured calls in, rear-ended, all I got was license plate:

“Claims Data Fill takes that license plate, reach out to DMV to get vehicle registration to get VIN number, we have policy database and get the carrier and policy information, take the registered owner, go out to public records, pull back their address, date of birth, telephone number, social security, wrap that into a package and put it back into our system, 88% of the time done in less than 5 seconds.

“Take minimum information provided at first notice of loss, provide a fraud score at the initial notice of loss. Daily monitoring of claim every time new information comes in, able to run various scores: fraud scores, severity score.”

Example: TransUnion Criminal History Score

“TransUnion recently evaluated the predictive power of court record violation data (including criminal and traffic violations)

“While a court record violation is created during the initial citation, the state MVR is updated later and may be delayed depending on a consumer’s response to the citation. For example, if someone pleads guilty and pays a ticket immediately, the state MVR will be updated in approximately two months. If the ticket is disputed in court, in contrast, the state MVR may not be updated for 6–19 months or longer.

“Also, as court records are created when the initial citation is issued, they provide insight into violations beyond those that ultimately end up on the MVR—such as violation dismissals, violation downgrades, and pre-adjudicated or open tickets.”

Predictive Analytics – Bias in Data

TransUnion Criminal History Score: Consider the disparities in criminal violations by race as illustrated in reports on policing in Ferguson, Missouri and Baltimore, Maryland.

Insurance Claim Fraud Scores: Consider the possibility that the foundational information of fraudulent claims – the target variable in the predictive model – may be biased for a variety of reasons, including historical scrutiny on claims submitted by minorities or biases of claims settlement personnel.

Big Data Algorithms Can Reflect and Perpetuate Historical Inequities

Barocas and Selbst: *Big Data's Disparate Impact*

Advocates of algorithmic techniques like data mining argue that they eliminate human biases from the decision-making process. But an algorithm is only as good as the data it works with. Data mining can inherit the prejudices of prior decision-makers or reflect the widespread biases that persist in society at large. Often, the "patterns" it discovers are simply preexisting societal patterns of inequality and exclusion. Unthinking reliance on data mining can deny members of vulnerable groups full participation in society.

New York Times, August 10, 2015: Algorithms and Bias: Q. and A. With Cynthia Dwork

> Algorithms have become one of the most powerful arbiters in our lives. They make decisions about the news we read, the jobs we get, the people we meet, the schools we attend and the ads we see. Yet there is growing evidence that algorithms and other types of software can discriminate. The people who write them incorporate their biases, and algorithms often learn from human behavior, so they reflect the biases we hold.
>
> Q: Some people have argued that algorithms eliminate discrimination because they make decisions based on data, free of human bias. Others say algorithms reflect and perpetuate human biases. What do you think?
>
> A: Algorithms do not automatically eliminate bias. . . .Historical biases in the . . .data will be learned by the algorithm, and past discrimination will lead to future discrimination.

Fairness means that similar people are treated similarly. ***A true understanding of who should be considered similar for a particular classification task requires knowledge of sensitive attributes, and removing those attributes from consideration can introduce unfairness and harm utility.***

Q: Should computer science education include lessons on how to be aware of these issues and the various approaches to addressing them?
A: Absolutely! First, students should learn that design choices in algorithms embody value judgments and therefore bias the way systems operate. They should also learn that these things are subtle: For example, designing an algorithm for targeted advertising that is gender neutral is more complicated than simply ensuring that gender is ignored. They need to understand that classification rules obtained by machine learning are not immune from bias, especially when historical data incorporates bias.

White House Report on Big Data

For all of these reasons, the civil rights community is concerned that such algorithmic decisions raise the specter of “redlining” in the digital economy—the potential to discriminate against the most vulnerable classes of our society under the guise of neutral algorithms. . . . .But the ability to segment the population and to stratify consumer experiences so seamlessly as to be almost undetectable demands greater review, especially when it comes to the practice of differential pricing and other potentially discriminatory practices. It will also be important to examine how algorithmically-driven decisions might exacerbate existing socio-economic disparities beyond the pricing of goods and services, including in education and workforce settings

Predictive Analytics in Insurance: Regulatory Action Needed

1. There has been a revolution in insurance pricing, marketing and claims settlement resulting from insurers' use of Big Data -- massive databases of new insurance and non-insurance, personal consumer information with associated data mining and predictive analytics and scoring.

2. Insurers' use of Big Data has huge potential to benefit consumers and insurers by transforming the insurer-consumer relationship and by discovering new insights into loss mitigation.

3. Insurers' use of Big Data has huge implications for fairness and affordability of insurance and for regulators' ability to keep up with the changes and protect consumers from unfair practices

4. Big Data has massively increased the market power of insurers versus consumers and versus regulators. The balance of knowledge between insurers and consumers has grown sharply in favor of insurers. Insurers' use of 21st century data and analytics is out of balance with regulators' 20th century authorities and tools.

5. Market forces alone – "free-market competition" – cannot and will not protect consumers from unfair insurer practices. So-called "innovation" without some consumer protection and public policy guardrails will lead to unfair outcomes.

6. Oversight and limited regulatory intervention can promote more competitive markets and faster adoption of innovative technologies that benefits consumers and fulfill public policy goals.