A Brief Guide to Legal Issues for Data-miners

Aaron Lai¹, Thomas Ho², and Raymond Wan³
¹ATM/eCommerce, Bank of America, Concord, CA, USA
²Home Equity Group, Wells Fargo Bank, San Francisco, CA USA
³InfoAtlas, Inc., Oakland, CA, USA

Abstract - Any competent data-miner should have already acquired strong technical skills. However, an outstanding data-miner will also understand the legal aspects of data analysis. It is because empirical analysis in the business world requires not only quantitative mastery but also strong business sense. This paper will provide an overview on several legal issues (consumer credit, investment strategy, and privacy) surrounding data mining for financial institutions, brokerage firms, and pharmaceutical companies. We are not trying to write a comprehensive guide for data mining practitioners; the purpose of this paper is to make the readers calling for their compliance and legal departments of sound direction, because, “tomorrow is a busy day”.

Keywords: data mining, legal, financial institutions, pharmaceutical, consumer

1 Introduction

This paper is divided into four sections. The first section is Consumer Credit. Consumer lending policy affects every citizen and so it is heavily regulated by the government to ensure fair access to all. In this section, we will provide general introduction to the major legislations in this area and explain how an innocent data-miner may sidestep on legal traps. The second section is Privacy. In 2001, US Congress enacted the Gramm-Leach-Bliley Act (GLBA) to protect the non-public financial information of consumer. GLBA has imposed a new barrier of information sharing and thus it created new challenges for data analysis. We will discuss the GLBA and other privacy laws, including the USA Patriot Act 2001. The third section is Fiduciary Duties. Investment companies are bounded by various laws to protect clients' interest. Lastly, we will discuss the implication of HIPAA (The Health Insurance Portability and Accountability Act).

Of course, we are not lawyers and we are not providing any legal advice here. Please contact your compliance/legal department for legal opinions. Everything expressed or implied in this paper is our own personal opinions and they do not necessarily reflect those of our employers or affiliates.

2 Legal Issues for Financial Institutions

2.1 Consumer Credit

Financial institutions that provide any form of credits such as loan or installment to consumer are governed by many laws. Government has actively updated and enforced the law. It is very costly to ignore them as the punishment for violation could be in million dollars!

2.1.1 Fair Credit Reporting Act (FCRA)

“The Fair Credit Reporting Act (FCRA), enforced by the Federal Trade Commission, is designed to promote accuracy and ensure the privacy of the information used in consumer report.” Since most of the reporting responsibilities are rest on the credit bureaus (Experian, Equifax and Trans-Union), it is mainly an operation and compliance issue for other financial institutions. However, analysts could still fall into some traps. For example, FCRA states that all seven years or older adverse credit information could not be used in credit decision. Consider the following situation: an analyst had done an analysis on the historical data from the datamart of a financial institution. He/she found that customers who had a 90-day delinquency within ten years tended to have future missing payments. Although this could be a valid finding, it would be illegal to use that information.

2.1.2 Equal Credit Opportunity Act (ECOA)

Equal Credit Opportunity Act (ECOA) is the most common regulation faced by financial institutions. Here below is an extract:

The Equal Credit Opportunity Act (ECOA), 15 U.S.C. 1691 et seq. prohibits creditors from discriminating against credit applicants on the basis of race, color, religion, national origin, sex, marital status, age, or because an applicant receives income from a public assistance program.

The basic idea of ECOA is to protect the consumer from being discriminated by creditors. Even though the criteria
are quite straightforward and most people understand it, analysts could still have unwillingly violated it.

One of the biggest problems of using decision support system (DSS) to perform credit scoring is the “black-box” approach. The company will have difficulties to defend a credit decision in court if the software “unintentionally” discriminated against some group of people. For example, if the neural network software found that marital status was correlated to income (married couples often had jointed income and thus were belonged higher income groups), the software might think that married couples were better borrowers. The user of the software might not aware of this and the computer programmer might not know of the ECOA requirement. It could cost the company a lot of money when lawsuits were initiated.

Even statistical modeling or explicit analyses will have problems as well. Years ago, a modeler used the social security number to build a credit decision model with logistic regression. The model performed well and the test results were satisfactory. However, the company could not use it. Why? It was because using social security number as a predictive variable would implicitly violate ECOA. The first three digits of the social security number is the regional office that the person first registered. The company might unintentionally discriminate against recent immigrants, as they were usually concentrated in a few regional offices.

Many major financial institutions have developed Enterprise Prospect Database (EPD) projects to leverage the advancement of database technology and analytical techniques for marketing purpose. The idea is very simple: to build a database so that company can solicit prospects continuously.

In the past, the company needed to follow a model-building process similar to the diagram below. This was a long, laborious, and costly process. It took anywhere from three months to one year to complete the whole process.

According to the diagram below, EPD is an expensive project — a major bank would need to spend over $20 million to do it!

2.1.3 Fair Debt Collection Practices Act (FDCPA)

As stated in FDCPA [1], the purpose of the Act is “to eliminate abusive debt collection practices by debt collectors, to insure that those debt collectors who refrain from using abusive debt collection practices are not competitively disadvantaged, and to promote consistent State action to protect consumers against debt collection abuses.” This act covers many debts such as “Personal, family, and household debts are covered under the Act. This includes money owed for the purpose of an automobile, for medical care, or for charge accounts.” There are many restrictions on what debt collectors could or could not do in collection. The implication of this act for data-miners is that some of the results cannot be used because their conclusions might violate the FDCPA. An example of this is that the collectors cannot contact the consumer by postcard even though they could save money.

2.1.4 Fair Credit Billing Act (FCBA)

The purpose of this Act as stated in FCBA [2] is “to protect the consumer against inaccurate and unfair credit billing and credit card practices.” Many people think that this is basically a compliance issue on operation procedures. Unfortunately, this is not true. For example, FCBA [3] has stated some conditions where cash discount is not appropriate even the marketers want to use this in promotion.

2.1.5 Truth in Lending Act (TILA)

The FTC has stated that “The main purpose of the Truth in Lending Act is to assure the meaning disclosure of consumer credit and lease terms, including those in
advertisements, so that consumers can easily compare terms and shop wisely for credit.”

It is a complicated law that spans across many different areas. It governs the disclosure information (e.g. point size of APR disclosure) and also the display of advertising information. Even though marketers often use some psychological tricks such as mental accounting to get better business result in marketing other products or services. They might not be able to do that in consumer loans or consumer leases because those practices might violate of TILA.

2.1.6 OCC Regulations and Other Regulators

Office of Comptroller of the Currency (OCC) charters, regulates, and supervises all national banks. It also supervises the federal branches and agencies of foreign banks. One of the most important regulations for direct marketers is Community Reinvestment Act (CRA). This act is enacted to ensure that low and moderate income families have access to credit. Many direct marketers have used spatial database or GIS (Geographic Information System) to analyze the geographical distribution of prospects so that they could identify the most profitable areas/segments. Even though this is a common practice for direct marketers to not solicit some geographical areas, or “red-lining”, it is a blunt violation of CRA. A rule-of-thumb in doing spatial analysis is not to use more than the first three ZIP code digits. Credit product solicitation and decision could not be based on geographical area because doing so would create unfair treatment to low and moderate income residential areas.

2.2 Privacy Issues

2.2.1 Gramm-Leach-Bliley Act

Privacy is an important issue in our electronic age. The Gramm-Leach-Bliley Act (GLBA) is aimed at protecting the non-public personal information of consumer. This is also a law enforced not just FTC.

In a nutshell, GLBA requires financial institutions to properly disclose their company policy on the use of Nonpublic Personal Information (NPI) and allow the customers to disallow (“opt-out”) the companies to use or disclose that information. If a holding financial institution has more than one line of business and they are separate legal entities, GLBA might prohibit them sharing customer information. It is because “The term “affiliate” means any company that controls, is controlled by, or is under common control with another company.” This law requires companies to keep their suppression list (for those opted-out customers) up-to-date or else they might solicit “opt-out” customers. The IS/IT department should well aware of these requirements and makes sure that the prospect database is in compliance with the law.

“Shares of wallet” analysis is very common in banking. The bank wants to know the number of relationships with each individual customer and household, including every product from every line of the business. Facilitating cross-sell and up-sell is the major driving forces of building an Enterprise Datawarehouse (EDW). With GLBA privacy law, analyst might not use some Nonpublic Personal Information (NPI) such as mortgage balance to solicit customer on product such as student loan. The system architect of EDW should aware of these regulatory constraints in building the enterprise-wide data-sharing process. On the other hand, the analysts should understand that even if some information is available and predictive, it might not be legal to use for data analysis and modeling.

2.2.2 Federal Trade Commission (FTC) on Standards for Safeguarding Customer Information

On September 7, 2000, FTC has released an “Advance Notice of Proposed Rulemaking” in Federal Register titled “Privacy of Customer Financial Information—Security”. According to the document, the statutory objectives of this rule are:

1. Anticipation of Threats or Hazards to Security or Integrity
2. Preventing Unwarranted Access and Use
3. Insuring Security and Confidentiality
4. Consideration of Other Agencies’ Safeguards Standards

Information security will be the main issue for these rules and it is extended beyond typical computer security issues. The financial institution should not only consider the computer security issues such as protection from unauthorized usage, but should also reevaluate company policies that govern internal staffs, as well as external vendors that act as agents on behalf of the company in providing services to the customer or accessing customer information.

2.2.3 Unsolicited Commercial Email (UCE) Act

The UCE Act is also called “antispam” law because it is targeted at the “junk” emails all of us receive everyday. This is a state level law and it is up to each state to enact and enforce it. As a result, each state has its own version of UCE and discrepancies could be found. The basic idea of UCE is that email marketers are required to notice the customers for permission before any marketing activities and getting the consent of the customers. [4]
For nationwide marketers, it is advisable to use the most restrictive standard, as it will be in compliance with all states. As a side note, anti-spam software is the latest challenge to data mining and artificial intelligence.

### 2.2.4 USA Patriot Act 2001

After September 11, national security became a top priority. President Bush signed a new law called “Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism (USA PATRIOT ACT) Act of 2001”. This is a legal extension of many existing laws; some of the sections would be expired after a certain period of time. For a quick synopsis of the law, you could find it here [5]. Since the law allows the government to have a greater freedom and jurisdiction in collection and using information from all sources, including those previously unlawful sources, it has created a strong controversy among civil right groups and online watchdog such as Electronic Frontier Foundation (EFF). Some of the sections which are particularly relevant to financial institutions are in Title III—International Money Laundering Abatement and Anti-Terrorist Financing Act of 2001, these includes:

- **Subtitle A**—International Counter Money Laundering and Related Measures
- **Subtitle B**—Bank Secrecy Act Amendments and Related Improvements
- **Subtitle C**—Currency Crimes and Protection

The whole document consists of 342 pages; it contains many complicated references and amendments. Financial institutions need to aware of the additional requirements on preventing money laundering and it is an amendment to Bank Secrecy Act.

USA Patriot Act had created a boom for data mining and data processing industry. First of all, transaction data is notoriously difficult to manage and analyze because they are voluminous; the company needs to store all of the data and match it against a known list. Moreover, most of the matching are not straightforward (e.g. “Apple” is not equal to “apple “) — it will involve certain level of fuzzy-matching. The companies are also required to develop some exception reports for suspicious money-laundering activities. It opened a new demand for unusual activity detection software.

### 2.2.5 Telemarketing Rules of FTC

FTC also has released another proposal on telemarketing practice. In January 2002, FTC has released a proposal to create a national “Do Not Call” registry. Under the current Telemarketing Sales Rule, it is illegal for the marketer to call a customer if he/she has already asked for “opt-out”. As stated in [6], the rule also:

- Restricts calling times to the hours between 8 a.m. and 9 p.m.
- Requires telemarketers to tell you it’s a sales call and who’s doing the selling before they make their pitch.
- Prohibits telemarketers from lying or misrepresenting any information.

The proposed rules would:

- Create a toll-free number for the consumers to place their number on a national “do not call” registry. Once the number is on the list, it would be illegal for a telemarketer to call it.
- Telemarketers would be required to “scrub” their lists and remove everyone on the “do not call” registry.

If these rules were enacted, the companies would need to maintain an up-to-date calling list and revise their telemarketing practices. All of the existing scoring models for outbound call should be revisited as well.

### 2.2.6 Social Security Number Law

On January 1, 2008, newly enacted Social Security Numbers laws took place in New York, Rhode Island, Tennessee and other 30 states. Those states have enacted laws that restrict or prohibit the collection, use, and disclosure of consumers Social Security Numbers. The New York law being the most restrictive definitions of Social Security Numbers and it extends the definition of Social Security Numbers in full as well as any number derived from the number – partial or scrambled. Before the enactment of the New York Social Security Numbers law, the restrictions are only for full Social Security Numbers or the last four digits.

The permissible uses of Social Security Numbers are now limited to collected, used, or released as required by state or federal law or court order, consumer reporting agency pursuant to FCRA, and investigation or prevention of fraud. Any public display, communication to unauthorized parties, or transmission over the Internet is prohibited.

This law could have significant impact in the merge/purge process of analysis or list-pulling exercise. It is a common practice for vendors to transmit 8-digit SSN and 5-digit ZIP as a combined matching key for direct mail respondents. With the enactment of this law, we could have lost the only way to identify responses!
Fiduciary Duty and Investment Recommendations

3.1 NASD General Rule on Suitability (Rule 2310) and NASD Regulation on Online Suitability

NASD (National Association of Securities Dealers, Inc.) has a Conduct Rule 2310 that governs the suitability of investment recommendations to the customers. The section (a) and (b) of it states that:

(a) In recommending to a customer the purchase, sale or exchange of any security, a member shall have reasonable grounds for believing that the recommendation is suitable for such customer upon the basis of the facts, if any, disclosed by such customer as to his other security holdings and as to his financial situation and needs.

(b) Prior to the execution of a transaction recommended to a non-institutional customer, other than transactions with customers where investments are limited to money market mutual funds, a member shall make reasonable efforts to obtain information concerning:

1. The customer’s financial status;
2. The customer’s tax status;
3. The customer’s investment objectives; and
4. Such other information used or considered to be reasonable by such member or registered representative in making recommendations to the customer.

The rule has also discussed how a firm should deal with its customer fairly and make investment recommendation only to suitable customer.

Due to the dramatic increase in online trading and web-based financial information portal, the NASD has published a policy statement on Online Suitability in April 2001. This is an extension of the NASD general suitability rule (Rule 2310) in the Internet world.

As stated in NASD [7], the suitability rule is:

The NASD’s suitability rule states that in recommending to a customer the purchase, sale, or exchange of any security, a member shall have reasonable grounds for believing that the recommendation is suitable for such customer. As the rule states, a member’s suitability obligation applies to securities that the member “recommends” to a customer. The NASD’s suitability rule generally has been violated when a broker/dealer “recommends” a security to a customer that might be suitable for that particular customer.

The key issue of this rule is the term “recommending”. Generally speaking, if a firm suggests a security to the customer and this action will result in a “call to action”, it is considered as a recommendation. The policy statement has explained under what circumstances a communication will be considered as a recommendation.

One of the implications of this online suitability rule is the use of data mining technology to suggest personalized information on a particular security to the customer. For example, if we use web tracking technology on customer preference to provide specific investment suggestions to that customer based on his/her web surfing pattern, then this action will be considered as a recommendation and it violates the rule. The business managers and data-miners of brokerage firms should be aware of this restriction in developing new web product.

3.2 U.S. Investment Company Act 1940 and U.S. Investment Advisers Act 1940

Both of the Acts are aimed at protecting the money of investing public from the misbehavior of investment communities. After the meltdown of financial market and the great depression, the government has taken action in regulating the financial markets by establishing a new government agency called Securities and Exchange Commission (SEC). These two Acts defined the fiduciary duties (i.e. treat it as if you own it) of the investment companies and investment advisors to the investors and what steps they have to take to comply with the law. The Acts also require proper disclosure on company information and investment policies to the investors.

Data-miners for investment data should aware that a theoretically or empirically sound investment approach you have found based on historical data might not be prudence in actual investment policy. For example, the mean-variance portfolio construction is standard in financial economic literature. However, it is not well established in the law of court nor it should be used as the single criterion in diversification. If the empirical result shows that two stocks are negatively correlated with another, in theory, we could say that we could diversify by buying two stocks together even if they belong to the same sector. It might not be a prudent suggestion as there is no guarantee that the negative correlation will persist into the future and we should diversify by investing in more than one sector.
3.3 Employee Retirement Income Security Act of 1974 (ERISA)

ERISA is enacted to protect the pension plan from "inappropriate" investment. The plan trustee is supposed to act solely at the interest of the plan participant and not the plan sponsors'. Take this principle to an extreme; if tobacco stocks were good investment, the trustees should not avoid investing in tobacco stocks even if the trust is for a lung cancer research company. In formulating the investment strategy for pension funds, the analyst should consider not only the investment value, but also the characteristics of the plan practitioners. In this aspect, it would be helpful to have input from actuaries and pension fund expert.

3.4 Uniform Management of Institutional Funds Act (UMIFA)

The UMIFA was resulted from the lack of laws in governing endowment and investment funds. The goal of this Act is, as stated in UMIFA [8], is:

1. A standard of prudent use of appreciation in invested funds;
2. Specific investment authority;
3. Authority to delegate investment decisions;
4. A standard of business care and prudence to guide governing boards in the exercise of their duties under the Act; and
5. A method of releasing restrictions on use of funds or selection of investments by donor acquiescence or court action.

This Act governs how an institutional fund should be managed and what would be a prudent action on investment due to the special nature of this kind of fund (e.g. use of appreciation fund). The investment analyst should understand that some actions are appropriate for general investment but not for endowment funds.

4 HIPAA and Data Mining

HIPAA (The Health Insurance Portability and Accountability Act, signed into federal law in 1996) was designed to safeguard the privacy of personal medical data. The two key concepts here are privacy and security.

HIPAA provides a national standard for the protection of information relating to an individual's health. HIPAA's security standards were crafted to be technology-neutral so as to give users wide latitude in developing their information security systems and procedures. For example, encryption is no longer mandated and can be based on risk assessments by companies when health data are transmitted over the Internet. If a doctor sends an e-mail to another physician about a patient consultation, encryption may not be necessary. But if a large healthcare organization is sending large amount of transactions, then they might want to encrypt the data. Data standardization from HIPAA helps delivering "clean" data to data-miners.

To ensure compliance, integrated database can record "opt-in" (patients consent to receive marketing materials from healthcare organizations) and “opt-out” (patients do not consent to receive marketing material) information. The organizations are required to exclude “opt-out” records from purchased or shared databases. If appropriate steps are taken, marketing professionals can continue to segment and target specific healthcare market niches using data mining techniques.

Data-miners in medical and social sciences often run into problem with the privacy law when they try to accumulate large amount of data from clinical trials or other sources. Privacy law stipulated that data must be stripped of all fields such as name, address, date-of-birth, Social Security Number (SSN) … etc. that would enable an individual to be identified. However, the anonymity is not guaranteed even if the database does not contain information that could easily identifies a person. By knowing an individual's postal code and birth date, one could identify an individual's personal information in a supposedly anonymous public database with 69% accuracy. Knowing gender raised the accuracy to 87%! [9] We data-miners must be sensitive to those concerns when collecting and using data, or else we could expose ourselves to potential legal or regulatory risk.

Researchers often need access to medical records without actually accessing the personally-identifying information. However, it is not trivial to determine how identifiable an individual is from some particular information about him/her. For instance, genetic data presents a particular problem. SNP (single nucleotide polymorphism) data on one gene for one person cannot be used to identify that particular person. However, several thousands SNP from the same person will significantly increase the likelihood of being identified.. Genetic data are also extremely sensitive because they could identify risk factors for diseases and other medical problems. Although HIPAA does not cover companies offering DNA testing for genealogy, legislation covering the issue of privacy and confidentiality of genetic data has been addressed on a state-by-state basis.

Information gathered for clinical trials, in addition to HIPAA rules, are required under the Department of Health and Human Services (HHS) [10] or the Food and Drug Administration (FDA) [11] to take measures to protect personal health information from inappropriate use or disclosure. Moreover, in clinical research, physician-
investigators often perform dual roles to the subject/patient: as a treating physician and as a researcher. For the treating physician, duties of confidentiality have long been established under well-known legal and ethical standards. The Privacy Rule adds to these existing obligations. The HHS and FDA Protection of Human Subjects Regulations are concerned with the risks associated with participation in research. The HIPAA rule is concerned with the risk to the subject's privacy associated with the use and disclosure of the subject's protected health information.

5 Conclusion

To quote Winston Churchill, “This is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning.” There will be new data mining technology developed to help company making decision faster and better. Companies are still facing competitions from not only their traditional competitors but also from competitions where companies are in a totally different business (e.g. WalMart as bank and Microsoft as healthcare information provider). To strive and prosper in this hyper-competitive environment, corporations need to understand their core competence and use their strength to attack the weaknesses of the competitors. Human capitals continue to be the most precious assets of the company as technology could be duplicated and bought, but people will need time and culture to cultivate. Laws and regulations are constantly changing and thus analysts should keep their knowledge up-to-date. This article by itself will not do you any good, but the readers of this article will gain some knowledge and better prepared for tomorrow’s challenges.

6 References

[10] Department of Health and Human Services (HHS) 45 CFR part 46