Tweets as predictors for the stock market

Twitter can be a surprisingly valuable source of data when you're building a predictive model for the stock market. A paper by Johan Bollen and his colleagues, "Twitter Mood Predicts the Stock Market," summarizes an analysis of about ten million tweets (by about three million users), which were collected and used to predict the performance of the stock market, up to six days in advance.

The study aggregated tweets by date, and limited its scope to only those (three million) tweets that explicitly contained sentiment-related expressions such as "I feel, I am feeling, I'm feeling, I don't feel" were considered in the analysis. The researchers used two tools in this classic example of opinion mining:

- **Opinion Finder** is a sentiment-analysis tool developed by researchers at the University of Pittsburgh, Cornell University, and the University of Utah. The tool mines text and provides a value that reflects whether the mood discovered in the text is negative or positive.
- **GPOMS (Google Profile of Mood States)** is a sentiment analysis tool provided by Google. The tool can analyze a text and generate six mood values that could be associated with that text: calm, happy, alert, sure, vital, and kind.

The research followed this general sequence of steps:

1. By aggregating the collected tweets by date and tracking the seven values of the discovered moods, the study generated a *time series* — a sequence of data points taken in order over time. The purpose was to discover and represent public mood over time.
2. For comparison, the researchers downloaded the time series for the Dow Jones Industrial Average (DIJA) closing values (posted on Yahoo! Finance) for the period of time during which they collected the tweets.
3. The study correlated the two time series, using *Granger causality analysis* — a statistical analysis that evaluates whether a time series can be used to forecast another time series — to investigate the hypothesis that public mood values can be used as indicators to predict future DJIA value over the same time period.
4. The study used a Fuzzy Neural Network model (see Chapter 7) to test the hypothesis that including public mood values enhances the prediction accuracy of DJIA.

Although the research did not provide a complete predictive model, this preliminary correlation of public mood to stock-market performance identifies a quest worth pursuing.

Target store predicts pregnant women

In an unintentionally invasive instance, the Target store chain used predictive analytics on big data to predict which of its customers were likely to be pregnant. (Charles Duhigg, a reporter at *The New York Times,* initially covered this story.) Target collected data on some specific items that couples were buying, such as vitamins, unscented lotions, books on pregnancy, and maternity clothing.

Using that data, Target developed predictive models for pregnancy among its customers. The models scored the likelihood of a given customer to be pregnant.

Keep in mind that predictive analytics models don't rely on only one factor (such as purchasing patterns) to predict the likelihood of an event. Target probably did not rely on only one factor to make its predictions. Rather, the model looked at factors that included purchase patterns of pregnancy-related products, age, relationship status, and websites visited. Most important, the resulting predictions were based on events that happened over a period of time, not on isolated events. For instance, a couple buys vitamins at some point in time, a pregnancy-guide magazine at another point in time, hand towels at yet another time, and maternity clothes at a still different time. Further, the same couple could have visited websites related to pre-pregnancy, or could have visited websites to look for baby names or lessons for couples on how to cope with the first days of pregnancy. (This information could have been saved from search queries done by the couple.) Once Target identified potential customers as probably pregnant, it could then send specialized coupons for products such as lotion and diapers to those customers.

Details of the exact model that Target used to predict customer pregnancy are not available. One way to build such a model, however, is to use classification-based prediction. (Note that this is not the only possible way, and may not be the approach used at Target.) The general procedure would look like this:

1. Collect data about past, current, or potential customers, and their activities over time in cyberspace.

   

   You can use one of the predictive analytics tools mentioned in this book to connect to data sources such as social networks, micro-blogs, blogs, and healthcare websites. Or you can buy third-party research to provide you with data. Note that big data may have all kinds of emerging properties.

2. Collect transactional data from customers who actually purchase the products you're interested in, some of which are pregnancy-related.
3. Select training data that will be used to build your classification-based model, and set aside some of the past data to use in testing your model.
4. Test the model until it's validated and you're happy with the accuracy of its performance on historical data.
5. Deploy your model. As new incoming data for a given customer arrives, your model will classify that customer as either potentially pregnant or not.

Twitter-based predictors of earthquakes

Another astonishing use of predictive analytics is to detect earthquakes. Yes, earthquakes. Researchers Sakaki, Okazaki, and Matsuo from the University of Tokyo — situated in a region known for seismic activity — used postings on the Twitter microblog social network to detect an earthquake in real time. A summary of their 2010 research ("Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors") was published in the proceedings of the 2013 International Conference on World Wide Web.

The researchers' approach was to utilize Twitter users as sensor that can signal an event through tweets. Because Twitter users tend to tweet several times daily, the researchers could capture, analyze and categorize tweets in real time. They sought to predict the occurrence of earthquakes of Intensity three or more by monitoring those tweets. One result of the research was an earthquake-based Twitter monitoring system that sends e-mails to registered users to notify them of an earthquake in progress. Apparently the registered users of this system received notification much faster than from the announcements broadcasted by the Japan Meteorological Agency. The Twitter-based system was based on a simple idea:

1. The earthquake-detection application starts collecting tweets about an event that's happening in real time.
2. The collected tweets would be used to trace the exact location of the earthquake.

One problem: Tweets containing the word *earthquake* may or may not be about an actual earthquake. The data collected was originally focused on tweets consisting of words directly related to an earthquake event — for example, such phrases as "Earthquake!" or "Now it's shaking!" The problem was that the meanings of such words might depend on context. *Shaking* crops up in phrases such as "someone is shaking hands with my boss" — and even *earthquake* might mean a topic rather than an event (as in, "I am attending an earthquake conference"). For that matter, the verb tense of the tweet might refer to a past event (as in a phrase such as, "the earthquake yesterday was scary").

To cut through these ambiguities, the researchers developed a classification-based predictive model based on the Support Vector Machine (see Chapter 7).

- A tweet would be classified as positive or negative on the basis of a simple principle: A positive tweet is about an actual earthquake; a negative tweet is not.
- Each tweet was represented by using three groups of features.

  The number of words in the tweet and the position of the query word within the tweet.

  The keywords in the tweet.

The words that precede and follow a query word such as *earthquake* in the tweet.

- The model makes these assumptions:

  That a tweet classified as positive contains the tweeter's geographical location.

  That a tweeter who sends a positive tweet can be interpreted as a virtual sensor.

  That such a tweeter is actively tweeting about actual events that are taking place.

Twitter-based predictors of political campaign outcomes

In a relatively short time, political activity has saturated online social media — and vice versa — even at the higher levels of government. At the United States House of Representatives, for example, it's common to see congressional staffers in the House gallery, busily typing tweets into their BlackBerries while attending a session. Every senator and congressman or congresswoman has a Twitter page — and they (or their staffers) have to keep it active, so they tweet about everything happening inside the House.

Even so, some things never change: A successful political campaign still focuses on making its candidate popular enough to get elected. A winning candidate is the one who can make a lot of people aware of him or her — and (most importantly) get people talking positively about him or her. That's where politics and social media grab hold of each other.

An Indiana University study has shown a statistically significant relationship between Twitter data and U.S. election results. DiGrazia et al. published a paper titled "More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior". (An electronic copy is available at: `http://ssrn.com/abstract=2235423`.) The study found a correlation between the number of times a candidate for the House of Representatives was mentioned on Twitter in the months before an election and that candidate's performance in that election. The conclusion: The more a candidate is mentioned on Twitter, the better.

According to a *Washington Post* article, the sentiments expressed in tweets as reactions to the political events of the 2012 elections matched the balance of public opinion (as indicated by a random-sample survey) about 25 percent of the time.

As Nick Kolakowski relates in an article published online at Slashdot (`http://slashdot.org/`), a team at the Oxford Internet Institute, led by Mark Graham, investigated the relationship between Twitter and election results. They collected thirty million Tweets in October 2012, and counted how many tweets mentioned the two presidential candidates. They found that Obama was mentioned in 132,771 tweets; Romney was mentioned in 120,637 tweets. The Institute translated the count into projected percentages of the popular vote — 52.4 percent for Obama versus 47.6 percent for Romney. At least in terms of popular votes, those figures predicted Obama's victory.

However, a certain ambiguity tended to cloud the picture: The user who tweeted *about* a candidate might not vote *for* that candidate. One way to unveil the intention of such a Twitter user would be to apply sentiment analysis to the text of the tweet. In fact, Graham admitted that they should have analyzed the sentiments of the tweets. Clearly, sentiment analysis plays a major role in building a predictive analytics model for such situations.

So, if you're building a model that seeks to predict victory or defeat for the next prominent political candidate, here's a general approach:

1. Start by collecting a comprehensive training dataset that consists of data about past political campaigns, and present data about all current candidates.

   Data should be gathered from microblogs such as Twitter or Tumblr, and also from news articles, YouTube videos (include the number of views and viewer comments), and other sources.

2. Count the mentions of the candidates from your sources.
3. Use sentiment analysis to count the number of positive, negative, and neutral mentions for each candidate.
4. As you iterate through the development of your model, make sure your analysis includes other criteria that affect elections and voting.

   Such factors include scandals, candidates' interviews, debates, people's views as determined by opinion mining, candidates' visits to other countries, sentiment analysis on the candidates' spouses, and so on.

5. *Geocode* — record the geographical coordinates of — your criteria so you can predict by locations.

   One or more of the features you identify will have predicting power; those are the features that indicate whether a candidate won a past election. Such results are relevant to your model's predictions.

When your training data has been gathered and cleaned, then a suitable model can be based on classification. At this point, you can use Support Vector Machines, decision trees, or an ensemble model (see Chapter 7 for more on these algorithms) that would base predictions on a set of current criteria *and* past data for each of the political candidates. The idea is to score the results; the higher score should tell you whether the candidate in question will win.

Six Use Cases for Data Science and Predictive Analytics

SeattleDataGuy Follow
Jan 28, 2018
https://medium.com/coriers/7-use-cases-for-data-science-and-predictive-analytics-e3616e9331f9

Data science is a tool that has been applied to many problems in the modern workplace. Thanks to faster computing and cheaper storage we have been able to predict and calculate outcomes that would have taken several times more human hours to process. Insurance claims analysts can now utilize algorithms to help detect fraudulent behavior, retail salespeople can better tailor your experience both online and in store all thanks to data science. We have combined a few examples of real life projects we have worked on as well as a few other ideas we know other teams are working on to help inspire your team. Let us know if you need help figuring out your next data science project!

Predicting the Best Retail Location

One of the true factors of business success is "Location, Location, Location". You have probably seen this to be true when you see a spot that always has a new restaurant or store. For some reason, it just will never succeed. This forces businesses to think long and hard about where is the best location for their business. The answer is where your customers are when they think about your product. But where is that?

This example is actually being taken on by a few companies. One example is Buxtonco. Buxtonco is answering where should you open your next business with data! Their site exclaims:

"*That any retailer can achieve greater success and growth by understanding their customer and that there is a science behind identifying who that customer is, where potential customers live, and which customers are the most valuable"*

The concept is brilliant. Think Facebook geo-fencing in real life. By looking for where your customers may spend their time, and what they might be doing in certain locations the technology can help determine where it would be best to open your next business. Whether that be a coffee shop or a dress store. Data science and machine learning can occasionally seem limited to the internet. However, information provides power both online and in real life.

Predicting why patients are being readmitted

Being able to predict patient readmission can help hospitals reduce their costs as well as increase population health. Knowing who is likely to be readmitted can also help data scientist find the "why" behind specific populations being readmitted. This is not just important because of public health but also because the affordable care act reduces the amount of medicaid for claims when readmission occur prior to 30 days.

Hospitals around the country are melding multiple data sources beyond just typical claims data to get insight into what is causing readmission. One of the common approaches is researching ties between readmission and socioeconomic data points like income, addresses, crime rates, and air pollution.

Similar to the way marketers are targeting customers using machine learning and product recommendation systems that factor socioeconomic data points to tell how to sell to a customer. Hospitals are trying to better tailor their care to help their patients based off of how other similar patients have responded in the past.

Even a phone call at the right time after an operation has been shown to reduce the amount of readmission that occurs. Sometimes the reason patients are readmitted can have nothing to do with how the doctors treated them in the hospital but instead it could be that the patient didn't understand how to take their medication, or they didn't have anyone at their house to help take care of them. Thus, being able to figure out the why behind the readmission can in turn fix it. Once policy makers understand the why, it is much easier to develop better practices to approach each patient.

Detecting insurance fraud

Insurance fraud costs companies and the consumers (who are subjected to higher rates) tens of billions of dollars a year. To add to the problem, attempting to prove claims are fraudulent can in turn costs the companies more than the original cost of the claim itself.

This is why many companies have been turning to machine learning and predictive models to detect fraud. This helps pinpoint more claims that should be researched by human auditors. This method doesn't just reduce the costs of human hours, it also increases the opportunity to reclaim stolen dollars from fraudulent claims.

Once you have a fine tuned algorithm, the accuracy and rate at which your team processes fraudulent claims will increase dramatically.

Brick and Mortar Stores Predicting Product Needs and Prices Live As You Walk Into The Store

The concept of targeting a price for a specific customer is a tried and true method that many companies have implemented(even before we coined the term "data scientists"). If a salesman thought you were wearing an expensive suit, then they might offer you the same car they sold earlier that day at a higher price. In the same way, now the computer can quantify the best price to encourage a customer to make the decision to buy while also maximizing profits(Like Orbitz Did In 2012 For Mac Users "Oh, you like spending $1200 on your computer…well here is your plane ticket + a $100 upcharge").

This isn't even limited to e-commerce! Image if in life retail stores actually start using previous purchase history as soon as a customer walks into the door(like in the Minority Report).

Perhaps it's a Men's Warehouse or a Macy's, pick your store. They could meld that data with other information like your LinkedIn profile and Glassdoor salary estimates. Now they will know how much money you make and your buying habits, maybe even some notes from the previous salesman or saleswoman. All of this combined would allow them to better tailor an experience for you and other customers like you.

For customers who enjoy buying clothes and other products in person this could help provide a major competitive advantage for Men's Warehouse or other similar companies that already have a tendency to focus on the experience not just the sale(who knows, maybe that is why their stock has doubled in the last 6 months…probably not). Plus, then companies can better plan which sales person to partner with which customer. Maybe they can predict that a customer will respond better to the hard sell vs. the softer approach. All of this paired with a human could massively increase sales and customer satisfaction.

Managing IT service desks is a balance of having enough tech support professionals to minimize wait time and keep customer satisfaction at a high and keeping costs low by not having too many people working at one time.

Detecting Who To Call Fundraisers

As someone who has managed a fund-raiser, automation only takes things so far when it comes to donors. Certain donors may respond better to custom emails, or slightly different worded messages, maybe they respond better to a phone call. This is where data science and targeted messages and approaches can help.

Marketing departments are already implementing techniques like A/B testing to their websites and emails to help convince customers to buy a product. The concept of finding the right donors isn't really different at all.

The key is to start collecting data and managing it efficiently. We have been talking to a few non-profits, and although this use case is a possibility, most of them don't have the data stored in any form of data storage besides excel, or a basic data base. This makes it difficult to pull out these insights. This is why step one is to creating a data system that will provide insights in the future.

Predicting When a Patient Needs Behavioral Health Procedures Partnered With Their Physical Medical Procedures

One third of the population suffering from physical ailments also suffer from an accompanying mental health condition exacerbating the physical illness, reducing quality of life, and increasing medical costs. Some companies like Quartet are finding that if they help improve the mental health along with the physical health of their customers, it helps improve their overall health and reduce costs for the patients. Quartet is working on a collaborative health ecosystem by curating effective care teams and combining their expertise with data-driven insights.

We have also worked with insurance providers on similar projects where we helped them calculate the overall ROI of their new behavioral health plan that they had implemented to help deal with a specific physical pathology. It not only opened their eyes to the effects of their program, it also found 300k of savings. We are glad to see that larger companies like Quartet are taking this problem on as well!

Data science is a tool that allows companies to better serve their customer and their bottom line. However, it all starts with making sure your company is asking the right questions. If a company doesn't start with the right use cases and questions, it can cost thousands to millions of dollars. Most of this comes down to communication breakdowns. It can be very difficult to translate abstract business directives into concrete models and reports that provide the impact and influence on decision making that was required.

Our team wants to help equip your data scientists with the tools to increase their personal growth and your departments performance. If you want to start seeing growth in your team and your bottom line, then please feel free to contact us her