Predictive Modeling A Very Brief Introduction

A “too perfect” Interactive Discussion

Agenda

- "Too perfect" Board
  - INTRO
  - BASICS
  - A PASTA SAUCE EXAMPLE
  - A MONEY EXAMPLE (with formulas !)
- How to Play
  - You are expected to use your Smart Phones
  - There is a door prize when we reach the MONEY EXAMPLE (also for anyone who can guess why I have 4 categories of 7 questions each and why the values are $100 - $700. Hint : It's math related)

“Too Perfect” Game Board

INTRO	BASICS	PASTA SAUCE	MONEY
$100	$100	$100	$100
$200	$200	$200	$200
$300	$300	$300	$300
$400	$400	$400	$400
$500	$500	$500	$500
$600	$600	$600	$600
$700	$700	$700	$700

ZURICH

INTRO 100

- The decade the insurance industry begins to use predictive modeling.
  - Used in industries other than insurance since late 1970's
  - Started in Personal Lines in the late **1980**'s
  - Extended to Commercial Lines around 2005

INTRO 200

- Terminology
  - At its core, predictor variables (**X**) are functionally linked to an observed response (**Y**)
  - Example : Suppose we want to predict an individual's height as an adult
    - Y would be the height. It is often referred to as the Target Variable.
    - X could be the height at birth, the parent's height, other relatives' height, etc.
    - Are there any other factors that should be considered ? What do you think ?

ZURICH

INTRO 300

- The objective and result of a predictive model.
  - It is cutting edge, seeking to find and quantify previously hidden relationships
    - Who likes ice cream ?
    - Who likes getting packages ?

ZURICH

INTRO 400

- Types of variables in the final model
  - Lots (and lots) of **objective quantifiable variables** are considered; **the final cut reflects careful consideration and discussion with the end users.**
  - Insured Characteristics: e.g., number of vehicles, sprinkler system installed or not, cumulative loss experience in the past 3 years
  - Socio-demographic and/or geographic : e.g., total crime index , average annual precipitation

ZURICH

INTRO 500

- Types of models / (alternatively – things we can test for)
  - Policy level losses (**Pure Premium, Frequency, Severity, for example**)
  - Claims (Notice)
  - Elasticity and Probability of Retention (Logistic)
  - Number of subscribers to a new marketing campaign
  - Spending habits and credit card design
  - Sports
  - Pasta Sauce and $

INTRO 600

- A most critical consideration
  - **User Buy-In** is crucial to the success of any Predictive Model
    - The entire model must be explainable in layman's terms
      - Who has been to Alaska ?
      - How do you travel between Haines and Skagway ?
    - The model must pass reasonability tests, sensitivity analyses, etc.

INTRO 700

- Final commentary on variables
  - The variables must make sense but also be correlated with the target.
  - They should be objective, quantifiable, and relatively easily obtainable/updatable with minimal effort and cost.
  - Finally they should be reasonably consistent from year to year.

BASICS 100

- Broad sources of data
  - Internal
    - Exists within a company's data warehouses, generally at a very low level, example class code.
  - External
    - Exists outside of company warehouses, generally at much higher levels, such as state, zip code, county.

BASICS 200

- Examples of Internal data
  - Number of claims, size of loss
  - Number of states, diversity of risk (can be proxied by number of class codes)

BASICS 300

- Examples of External data
  - **Financial**
    - D & B credit score, number of bankruptcies, liens, payment history
  - **Demographic**
    - Relative pay by county, % of labor intensive workforce by county or ZIP
  - **Economic**
    - CPI, unemployment

BASICS 400

- This page left intentionally Blank
  - **I've always wanted to do this !**

ZURICH

BASICS 500

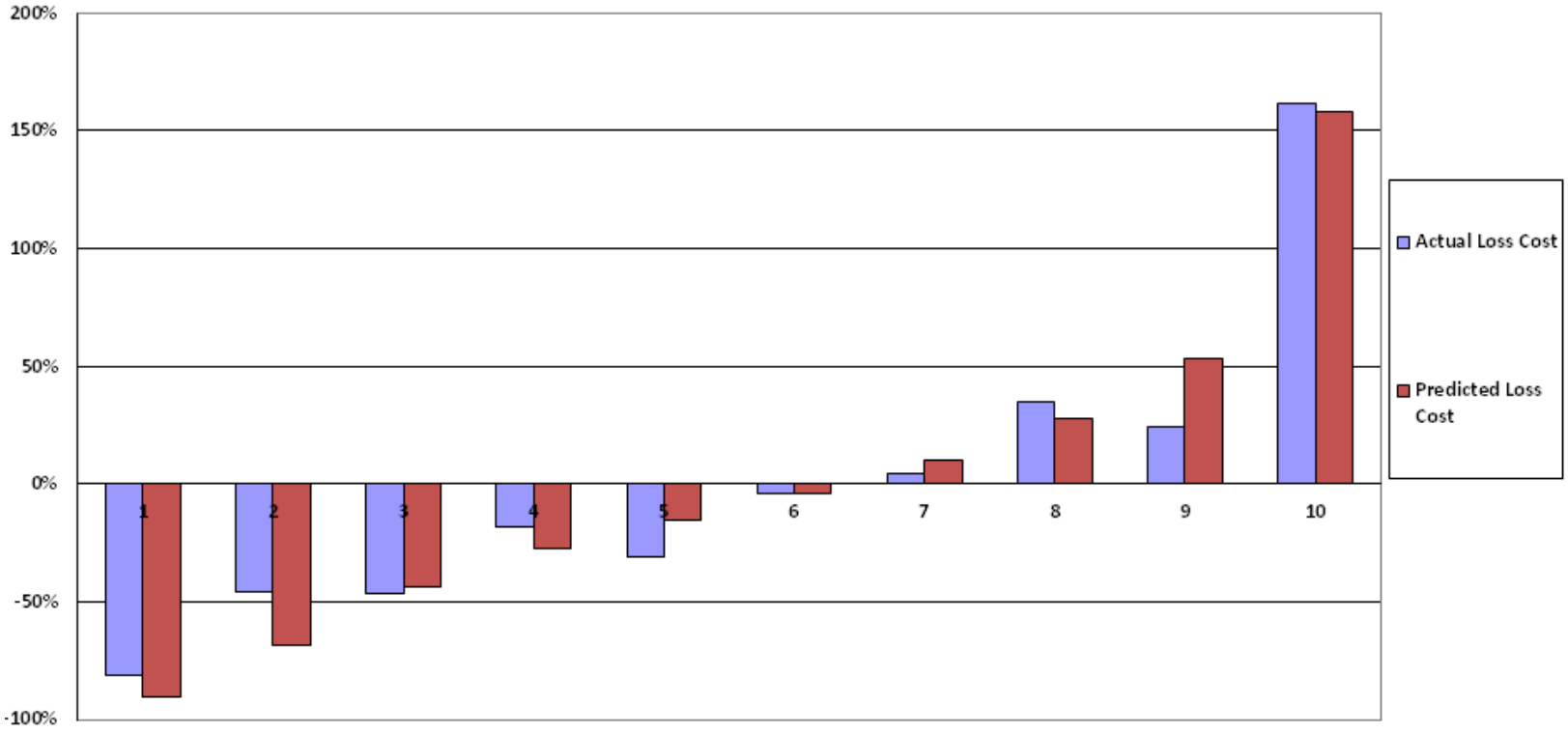
- Model performance and evaluation
  - **So the fitted values Y-hat are really, really close to the original Y's in aggregate**
    - Are we done ?
    - How do we assess the relative power of the model ?

BASICS 600

- Model performance and evaluation
  - One additional consideration should be parsimony
  - In other words, we want to avoid overfitting
  - Seek the fewest number of predictor variables that balance fit with segmentation

BASICS 700

- LIFT
  - Range of relativity between groups or bins
  - Generally speaking, we want a larger (wider) spread of the relativities between the lowest and highest bins on a predicted basis than actual while still maintaining a reasonable level of fit for each of the bins.



PASTA SAUCE 100

ZURICH

- The prediction of interest for Gaussian Gourmet.
  - The purveyor of premium pasta products (how's that for alliteration ?)
  - For a 0^{th} step, pre model design, what are some target variables we might be interested in predicting ?
    - Profits
    - Sales/revenue
    - Units produced
    - Distribution
      - Regional (do certain parts of the country tend to buy/eat more pasta products)
      - Daily (is more sold/eaten on certain days of the week and/or in certain areas)
    - **Taste rating for a sauce**

PASTA SAUCE 200

- The **S**auce **T**aste **I**ndex
  - **Our target variable** is STI, as we wish to predict this based on various input “variables”

PASTA SAUCE 300

- The motivating question for Gaussian Gourmet
  - We all know how to make a decent sauce, i.e. good ingredients coupled with proper cooking techniques
  - But what if we need to know how good our sauce will be if ingredients become expensive or scarce. What if the cost to run the cookers is increasing. **Are we using the right combination of ingredients; are we using too many ingredients ?**

PASTA SAUCE 400

- The range of score on the STI scale.
  - Range of scores on the STI scale **1 - 100**
  - 1 is not fit for human consumption; 100 sublime perfection.
  - 50 would be deemed average

PASTA SAUCE 500

- Variables
  - **Tomatoes (of course)**
  - Olive Oil
  - Spices
  - Time and cooking method ?

PASTA SAUCE 600

- Variables – a little more specific
  - **What kind of tomatoes**
    - **Roma, San Marzano, etc (what would be a corollary to the insurance world in terms of losses ?**
  - What kind of oil
    - Virgin, extra virgin
    - Italian, Greek, Spanish?
    - Q : do certain tomatoes and oils "interact" more/less favorably
  - Which spices
    - Basil, oregano (Mexican or Turkish ?), etc, etc, etc

PASTA SAUCE 700

- A few simple observed values

Observation	San Marzano	Roma	Beefsteak	Other_Tom	EVOO - I	EVOO - G	EVOO - S	VOO - I	VOO - G	VOO - S	PCT_SAN	STI
1	7	1	1	1	0	0	0	0	0	1	0.7	50
2	8	1	0	1	0	0	0	0	1	0	0.8	70
3	9	1	0	0	0	0	0	1	0	0	0.9	70
4	9	0	0	0	0	0	1	0	0	0	1	90
5	10	0	0	0	0	1	0	0	0	0	1	85
6	10	0	0	0	1	0	0	0	0	0	1	95

MONEY 100

- Boston, New York, Philadelphia, Cleveland, Richmond, Atlanta, Chicago, St. Louis, Minneapolis, Kansas City, Dallas, San Francisco
  - The 12 Federal Reserve Banks, in “letter order”
  - For even more fun, does anyone know the significance of
    - Dodgers, Cubs, Browns, Senators, Red Sox, Senators, Browns, Tigers, Browns, Tigers, Senators, Dodgers, Browns, Senators, Dodgers, Athletics, **Senators**, Yankees, Senators, Giants, Athletics, Senators, Athletics

MONEY 200

- We want to predict how much time it will take to collect at least 1 from each of the 12 Federal Reserve Banks ?
  - Everyone, take out your wallets and let's see how close we are

MONEY 300

- What are some variables we might consider
  - Your personal spending, as measured by the % of your paycheck that you save. The greater your spending, the faster you'll accumulate dollar bills in change and presumably the fewer overall dollar bills you'll need to collect .
  - Where you live, as measured by your home ZIP CODE. The closer you are to a Federal Reserve Bank, the more likely there will be bills from all banks (as they share) and the fewer you'll have to collect .
  - Additionally, if you happen to live in New York City, whose Federal Reserve Bank has 5 times the assets of the next bank (Richmond), the more likely you are to collect bills from all banks and the fewer you'll have to collect.

MONEY 400

- Let's build it (DOIT)
  - Model structure:
  - Constant +
  - {coefficient 1 * (% of paycheck saved)} +
  - {coefficient 2 * (ASSET_SIZE_RANK OF Federal Reserve Bank closest to you)} +
  - {coefficient 3 * (ln [distance between your home zip code and the Federal Reserve Bank closest to you)}

MONEY 500

- Example

Suppose you live in New York City, Zip Code 10009
Your closest Federal Reserve Bank is New York (B), which is located
in Zip Code 10045

Assume you save 25% of your paycheck

SOLVED

Coefficients	Variable Input	
-3.18	-3.18	
66.82	25.0%	You save 25% of your paycheck
0.97	1	Rank of the New York Federal Reserve Bank = 1
0.96	0.65752	Distance in miles = 1.93

16	Projected amount of time it takes to collect at least 1 bill from each of the 12 Federal Reserve Banks

MONEY 600

- What does it look like in matrix format
  - Raw Inputs in red, first 5 columns
  - Calculated inputs in **blue, RANK_CLOSEST_BANK and DIST_MIN**
  - 2 columns of calculated **blue** input required multiple additional columns of Data Prep including
    - Latitude and longitude for approximately 44,000 distinct Zip codes
    - an external variable of the Federal Reserve Banks assets and Zip code locations
  - Observation : I live in Zip code 21136, my closest Federal Reserve Bank is in Philadelphia, approximately 94 miles away (which is in hidden column DIST_C). DIST_A is how close I am to Boston, DIST_B is how close I am to New York

OBS	NAME	TIME	ZIPCODE	%_SAVED	RANK_CLOSEST_BANK	DIST_MIN	DIST_T	DIST_A	DIST_B
1	Jon Harbus	32.09004102	21136	45.0%	7	94.54395982	8454.967	360.9573	171.3193
2	Jon Harbus	18.38491934	21136	4.0%	7	94.54395982	8454.967	360.9573	171.3193
3	Jon Harbus	56.93394375	21136	69.0%	7	94.54395982	8454.967	360.9573	171.3193
4	Jon Harbus	23.223056	21136	24.0%	7	94.54395982	8454.967	360.9573	171.3193
5	Jon Harbus	44.1238064	21136	60.0%	7	94.54395982	8454.967	360.9573	171.3193
6	Jon Harbus	46.446112	21136	62.0%	7	94.54395982	8454.967	360.9573	171.3193
7	Jon Harbus	21.78953403	21136	19.0%	7	94.54395982	8454.967	360.9573	171.3193
8	Jon Harbus	55.15475801	21136	68.0%	7	94.54395982	8454.967	360.9573	171.3193

MONEY 700

- OK, how do we solve it
  - Depending on how we define *function* (i.e. how are the values of the response variable distributed) and how do the variables "work" together, (directly additive) the solution here can actually be done as a closed-form multiple regression via $(X^T X)^{-1} X^T Y$. Does this look familiar

OBS	NAME	TIME	%_SAVED	RANK_CLOSEST_BANK	DIST_MIN	FITTED VALUE	Const	X1	X2	X3	Error_(SQ)
1	Jon Harbus	32	0.45	7	4.55	38	-3.18	66.82	0.97	0.96	35
2	Jon Harbus	18	0.04	7	4.55	11	-3.18	66.82	0.97	0.96	60
3	Jon Harbus	57	0.69	7	4.55	54	-3.18	66.82	0.97	0.96	8
4	Jon Harbus	23	0.24	7	4.55	24	-3.18	66.82	0.97	0.96	1
5	Jon Harbus	44	0.6	7	4.55	48	-3.18	66.82	0.97	0.96	15
6	Jon Harbus	46	0.62	7	4.55	49	-3.18	66.82	0.97	0.96	9
7	Jon Harbus	22	0.19	7	4.55	21	-3.18	66.82	0.97	0.96	1
8	Jon Harbus	55	0.68	7	4.55	53	-3.18	66.82	0.97	0.96	3

FUN WITH MONEY – BONUS PSEUDO CODE

Our model is **TIME** = ***PCT_SAVED*** ***RANK_CLOSEST_BANK*** ***DIST_MIN***

To solve exactly as in $(X^T X)^{-1} X^T Y$ we would indicate the assumptions that the response variable is normally distributed and that the predictor variables undergo no transformation (Identity) as they are combined.

Something like this :

TIME = ***PCT_SAVED*** ***RANK_CLOSEST_BANK*** ***DIST_MIN \ none, Normal***

To use a different distribution function, we could use Poisson, gamma, etc.

We can also assume that the predictor variables are either inverted or exponentiated as they are combined.

Something like this :

TIME = ***PCT_SAVED*** ***RANK_CLOSEST_BANK*** ***DIST_MIN \ inverted, Poisson***

Thank you for participating

FEEDBACK / QUESTIONS

DISCLAIMER

The information in this presentation was compiled from sources believed to be reliable for informational purposes only. Any and all information contained herein is not intended to constitute advice (particularly not legal advice). Accordingly, persons requiring advice should consult independent advisors when developing programs and policies. We do not guarantee the accuracy of this information or any results and further assume no liability in connection with this presentation. We undertake no obligation to publicly update or revise any of this information, whether to reflect new information, future developments, events or circumstances or otherwise.

ATTRIBUTIONS

This presentation was created using PowerPoint® presentation manager. Microsoft Powerpoint is either a registered trademark or trademark of Microsoft Corporation in the United States and/or other countries

.