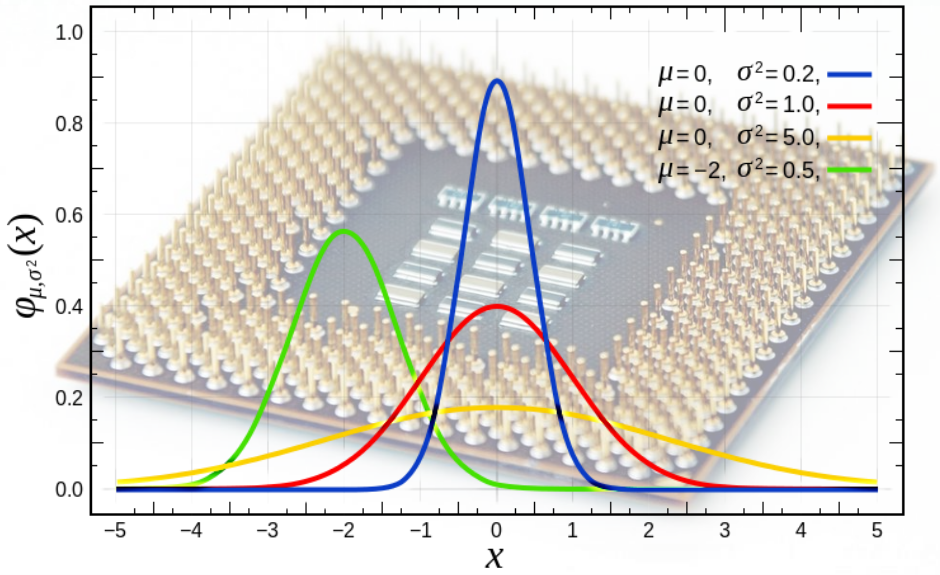
SMARTVISION Europe



Statistics vs Machine Learning
ASC 15th November 2018

Jarlath Quinn

Contents

- Machine Learning is Hot
- How did we get here?
- Techniques and Terms
- An Example
- Why Machine Learning matters
- Why Statistics *still* matters
- What can we expect in the future?

Machine Learning is Hot

Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil

Computer learns to detect skin ca more accurately than doctors

Artificial intelligence machine found 95% of melanomas in study compared to 86.6% for dermatologists

'Deep learning' – the hot topic in AI

Experts in the field are in demand and future managers would do well to grasp the concept

© Getty

Deloitte.

Machine learning: things are getting intense

Amazon uses data crunching skills to shape the future

It plays an ever greater role in life, but some uses of artificial intelligence are provoking concern

Machine-learning system could aid critical decisions in sepsis care

Model predicts whether ER patients suffering from sepsis urgently need a change in therapy.

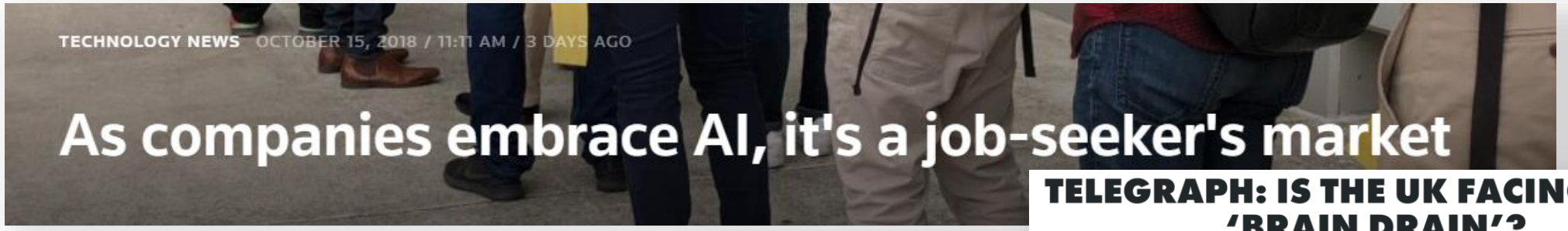
Harvard Business Review

EDUCATION

The Chairman of Nokia on Ensuring Every Employee Has a Basic Understanding of Machine Learning — Including Him

by Risto Siilasmaa

OCTOBER 04, 2018



TELEGRAPH: IS THE UK FACING AN AI 'BRAIN DRAIN'?

'A third of leading Machine Learning and AI specialists who have left the UK's top institutions are currently working at Silicon Valley tech firms' says the paper – as British start-ups just can't match the six-figure salaries on offer over the Pond

HOME » BREAKING NEWS » BUSINESS

Ireland is chasing to keep pace with data analytics revolution

Facebook Twitter Messenger LinkedIn WhatsApp More

Friday, November 09, 2018 - 10:21 AM

By Joe Dermod

7,069 views | Jun 25, 2018, 02:10am

The AI Skills Crisis And How To Close The Gap

Bernard Marr Cont

6 most in-demand IT positions in artificial intelligence

Machine learning engineer

A machine learning engineer is an advanced software developer. This job involves developing complex programs using predictive models. The goal of the machine learning engineer is to create programs that can learn and apply that knowledge in required situations without specific instructions.

Machine learning is a crucial part of AI and is already changing our lives. Machine learning engineers are needed to create most AI systems — from virtual personal assistants to self-driven cars, search engines, and more. If you are interested in becoming a machine learning engineer, you need to have excellent programming skills as well as a background in mathematics, and cloud technology.

October 19, 2018

Good News for Job Seekers With AI Skills: There is a Shortage of Talent

A short pool of AI-trained job seekers has slowed down hiring and impeded gro

By Stacy Stanford

TechRepublic. SEARCH Cloud CXO Software Start

CXO

Demand for AI talent exploding: Here are the 10 most in-demand jobs

Employer demand for AI-related roles has more than doubled over the past three years, according to Indeed.

By Alison DeNisco Rayome | March 1, 2018, 6:00 AM PST

TELEGRAPH: IS THE UK FACING AN AI 'BRAIN DRAIN'?

'A third of leading Machine Learning and AI specialists who have left the UK's top institutions are currently working at Silicon Valley tech firms' says the paper – as British start-ups just can't match the six-figure salaries on offer over the Pond

HOME » BREAKING NEWS » BUSINESS

Ireland is chasing to keep pace with data analytics revolution

Facebook Twitter Messenger LinkedIn WhatsApp More

Friday, November 09, 2018 - 10:21 AM

By Joe Dermo

7,069 views | Jun 25, 2018, 02:10am

The AI Skills Crisis And How To Close The Gap

6 most in-demand IT positions in artificial intelligence

Machine learning engineer

A machine learning engineer is an advanced software developer. This job involves developing complex programs using predictive models. The goal of the machine learning engineer is to create programs that can learn and apply that knowledge in required situations without specific instructions.

Machine learning is a crucial part of AI and is already changing our lives. Machine learning engineers are needed to create most AI systems — from virtual personal assistants to self-driven cars, search engines, and more. If you are interested in becoming a machine learning engineer, you need to have excellent programming skills as well as a background in mathematics, and cloud technology.

October 19, 2018

Good News for Job Seekers With AI Skills: There is a Shortage of Talent

A short pool of AI-trained job seekers has slowed down hiring and impeded gro

By Stacy Stanford

TechRepublic SEARCH Cloud CXO Software Start

CXO

Demand for AI talent exploding: Here are the 10 most in-demand jobs

Employer demand for AI-related roles has more than doubled over the past three years, according to Indeed.

By Alison DeNisco Rayome | March 1, 2018, 6:00 AM PST

Where is 'Statistics' in all this?

Let's compare search terms on a leading recruitment site

indeed

"Machine Learning"

Salary Estimate	
£35,000	(4266)
£45,000	(3418)
£50,000	(2920)
£60,000	(1803)
£70,000	(967)

"Statistics"

Salary Estimate	
£20,000	(4542)
£30,000	(3567)
£35,000	(3008)
£45,000	(1908)
£55,000	(1050)

How did we get here?

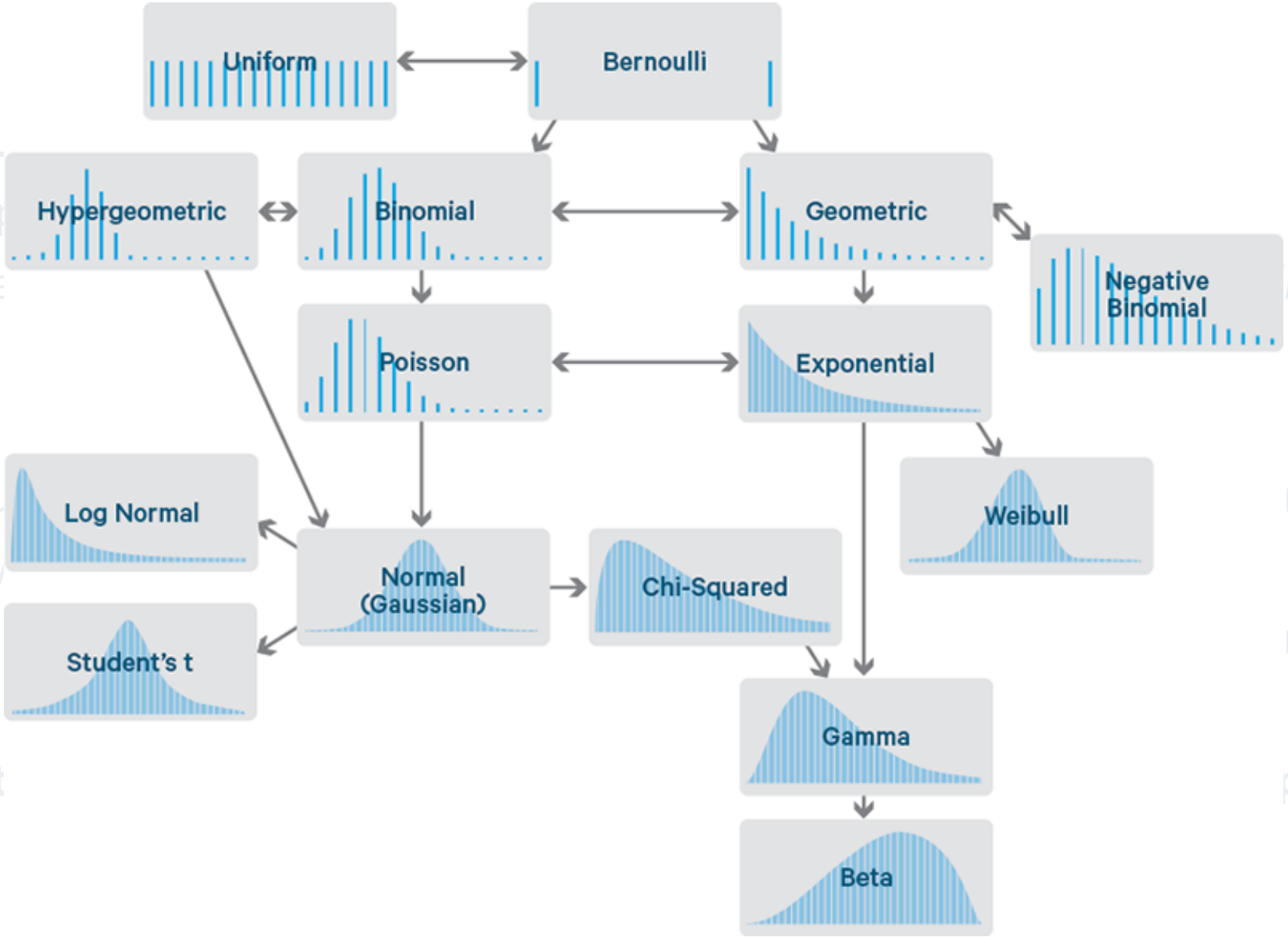
Timeline of Statistics and Machine Learning

- **17th Century**
  - Development of probability theory (Cardano, Pascal, Fermat)
  - John Gaunt - *Natural and Political Observations upon the Bills of Mortality* (1663)
- **18th Century** – Introduction of Bayes theorem
- **Late 19th and early 20th Century**
  - Introduction of standard deviation, correlation, linear regression (Galton, K. Pearson)
  - Null Hypothesis, Variance (Fisher)
  - Type II Error, Statistical Power, Confidence Intervals (E. Pearson, Neyman)

All of these 'tools' pre-date modern computing by utilising distributional approaches

- **17th Centu**
  - Develop
  - John Ga *ity* (1663)
- **18th Centu**
- **Late 19th a**
  - Introdu n, K. Pearson)
  - Null Hy
  - Type II n)

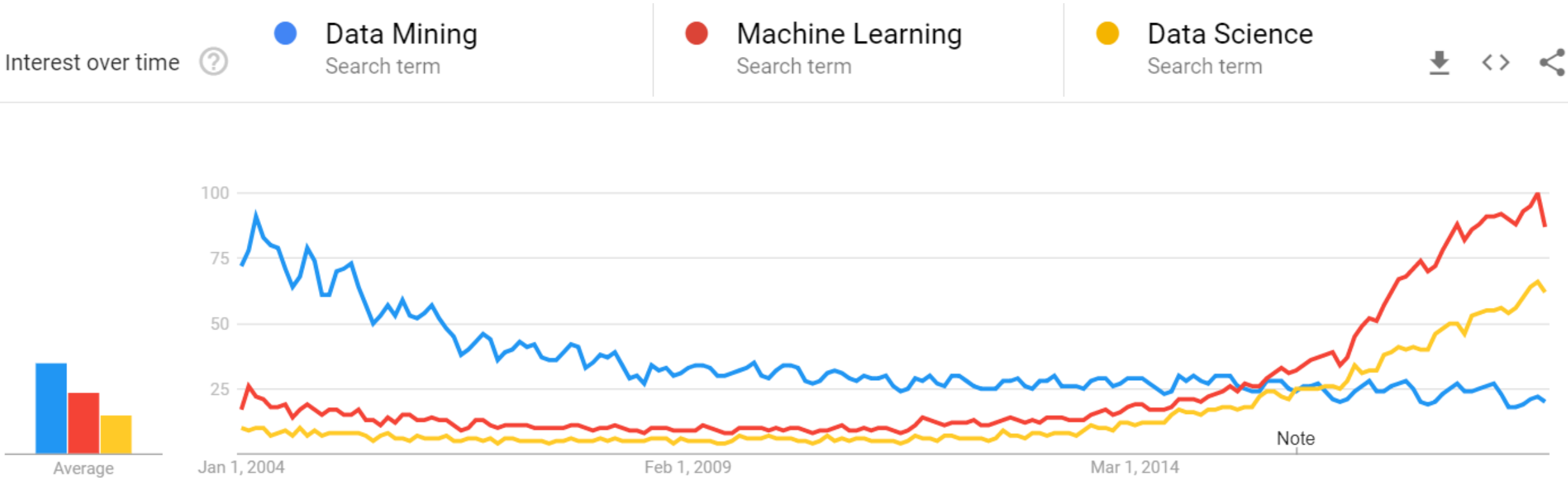
All of these 't proaches



Timeline of Statistics and Machine Learning

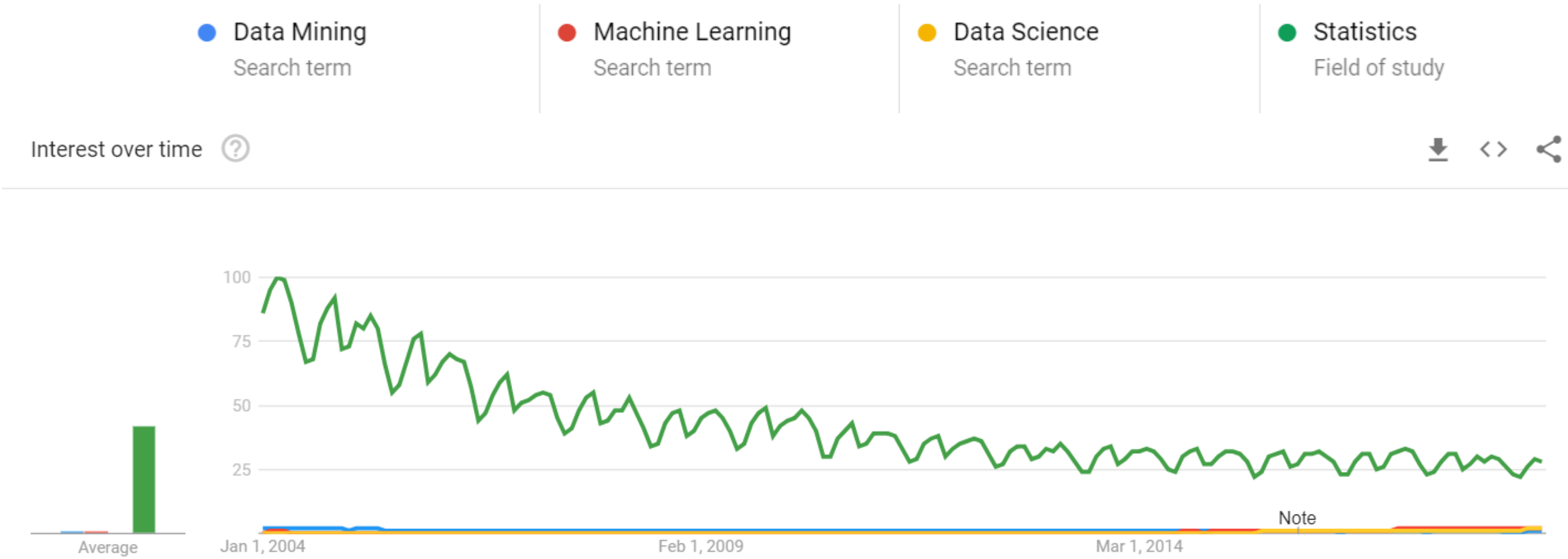
- **1951** – First neural network (Minsky & Edmonds) – the SNARC
- **1957** – Invention of the Perceptron (Rosenblatt)
- **1967** – Introduction of Nearest Neighbour algorithm
- **1970** – Introduction of Backpropagation (Linnainmaa)
- **1975** – Decision Tree algorithm (Quinlan) – ID3
- **1989** - Convolutional neural network used to read digits (LeCun)
- **1992** – Modern Support Vector Machines developed (Boser, Guyon & Vapnik)
- **1995** – Random Forest method proposed (Ho)
- **2003** – Adaptive Boosting (AdaBoost) wins Godel Prize (Freund, Schapire)
- **2011** – Alexnet Deep Learning network (Krizhevsky, Sutskever, Hinton)
- **2016** - Google's AlphaGo program beats professional human player

The New Terms on the Block




“My CPU is a neural-net processor; a learning computer.
The more contact I have with humans, the more I learn.”
SCHWARZENEGGER
THE TERMINATOR

But Statistics is still *very* important



Techniques and Terms

Statistics vs Machine Learning

Two analytical cultures

Stats Speak vs ML Speak

- Statistics
- Parameters
- Fitting
- Covariate
- Dummy coding
- Regression/Classification
- Density estimation, clustering
- Dependent
- Independent

- Machine Learning
- Weights
- Learning
- Feature
- One hot encoding
- Supervised Learning
- Unsupervised Learning
- Target
- Predictor

4 broad families of mining (Machine Learning) algorithms*

Type	Overview
Statistical	Often based on some form of regression – that originate in traditional statistics. Including: • **Linear Regression, Logistic Regression** • **Discriminant Analysis** • **Structured Equation Modelling/Latent Class (may be confirmatory)** • **Generalized Linear Models (GLM)** • Cluster/PCA/Factor
(Core) Machine Learning	Derived from research into Information Science and AI. Including: • **Neural - Multi-Layer Perceptron (MLP), Radial Basis Function (RBF)** • **Support Vector Machines (SVM)** • **Deep Learning** • Self organising maps
Rule Induction (/Decision Trees)	Rule induction algorithms use criteria from Statistics and Machine Learning to derive rules and trees that make predictions including: • **Classification and Regression Trees (CaRT)** • **CHAID** • **C5** • **Gradient Boosted Trees** • **Random Forests**
Association/Sequence	Find Associations and Sequences in the form … **IF A and B then C will happen with X% confidence** **CARMA, Apriori, SPADE, etc.**

I We used to say only this was I ML

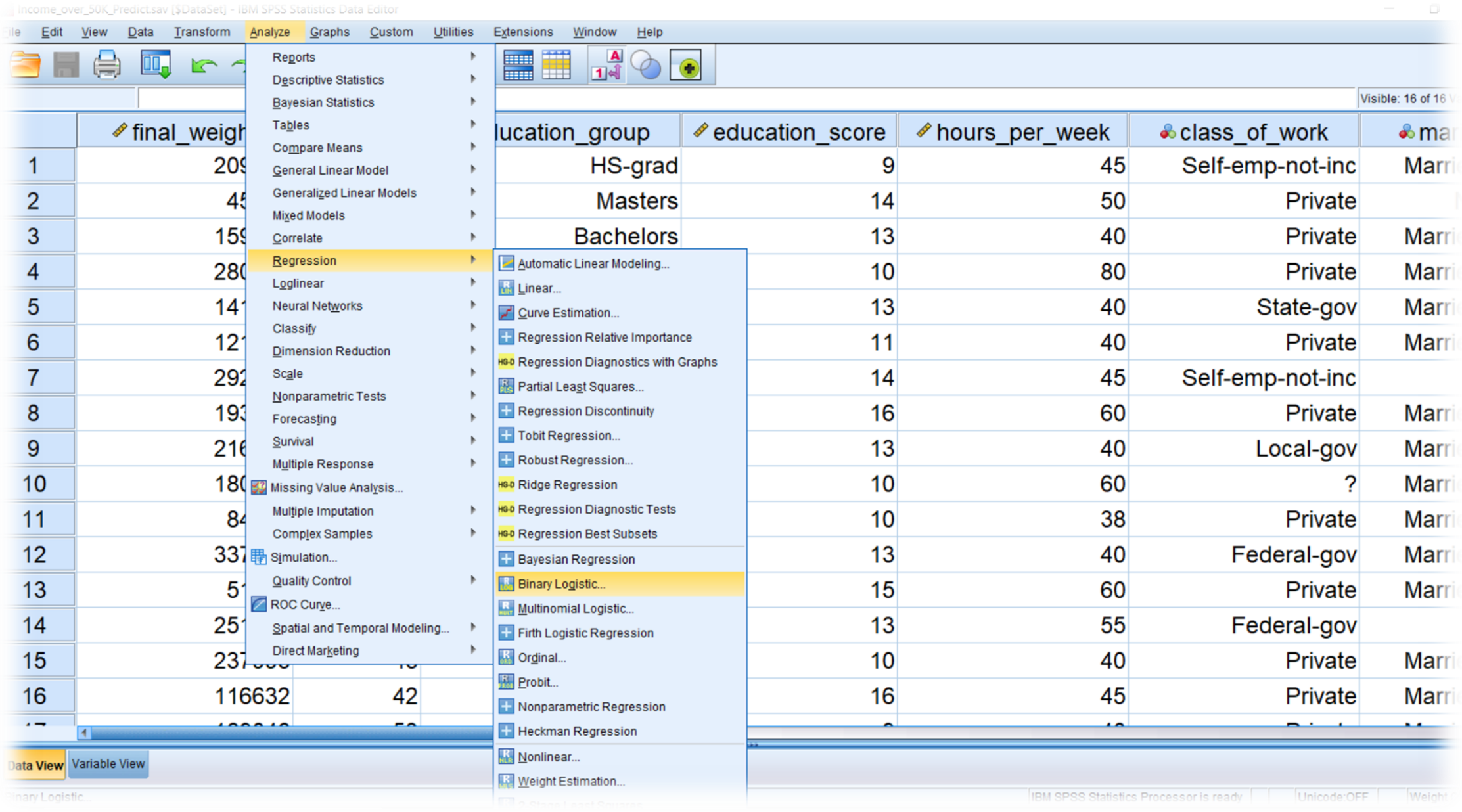
*Not an exhaustive list by any means

An Example

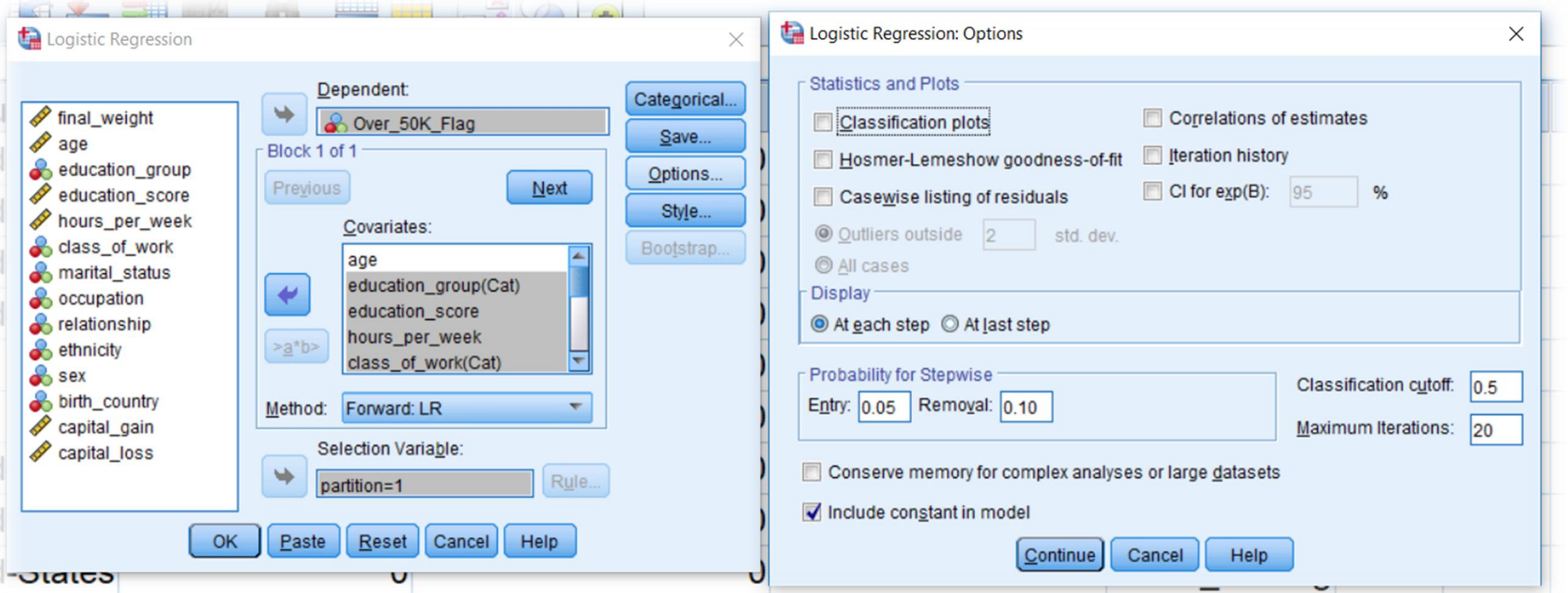
- **Data from 1994 US Census**
- **Contains range of demographic and employment related fields**
- **Goal is to predict whether respondent earns over $50K**
- **Data is randomly split (50/50) into separate ‘Training’ and Testing’ groups**

		Count	Column Percent
Over_50K_Flag	<=50K	1,645,823,488	76.6%
	>50K	504,079,916	23.4%
	Total	2,149,903,404	100.0%

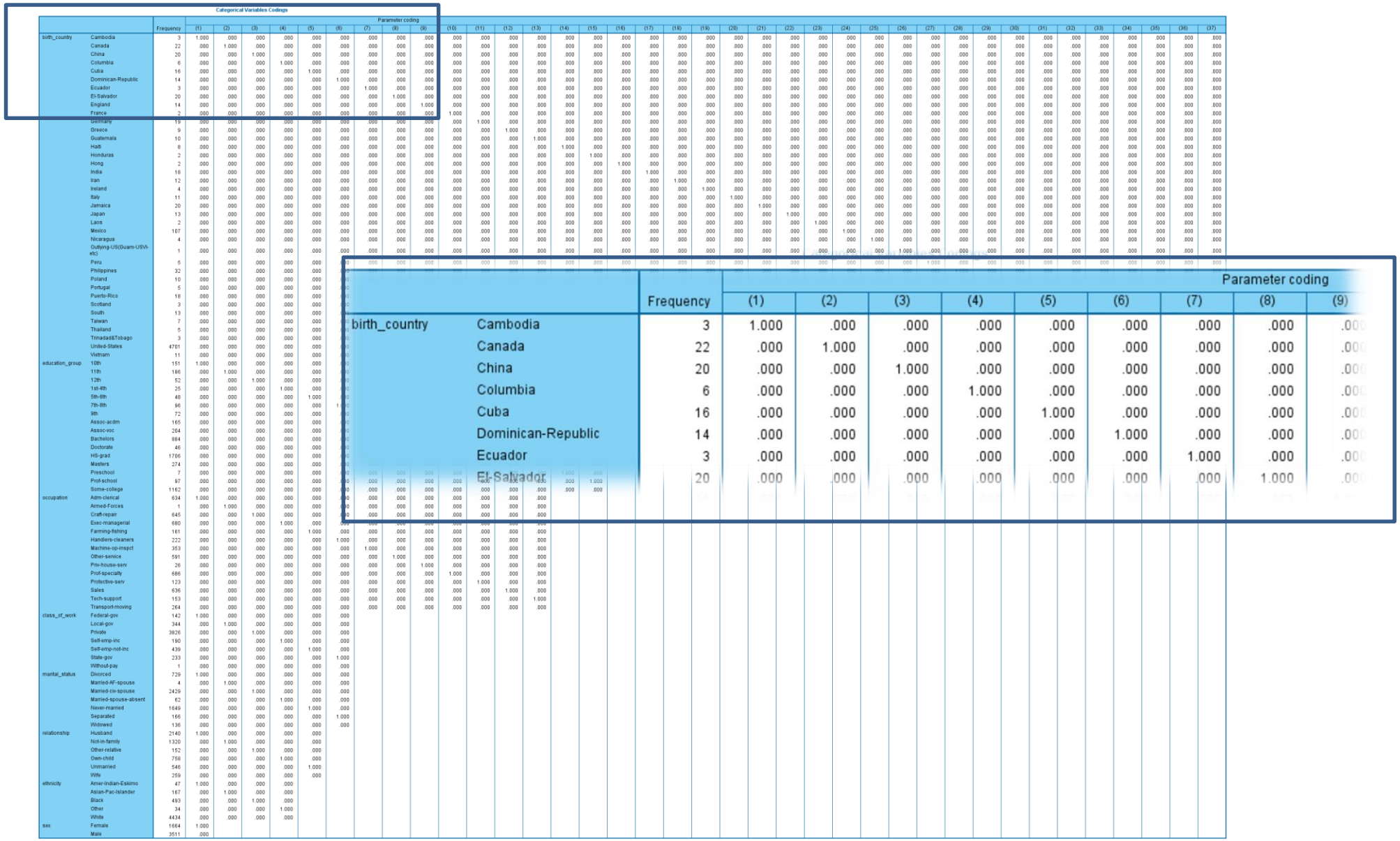
A Statistical Approach: Binary Logistic Regression



A Statistical Approach: Binary Logistic Regression – note the optional elements that can be *added* to the analysis output



A Statistical Approach: Binary Logistic Regression – First Output – Large table showing how categorical fields have been dummy coded



Categorical Variables Codings

		Frequency	Parameter coding (1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)	(31)	(32)	(33)	(34)	(35)	(36)	(37)
birth_country	Cambodia	3	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Canada	22	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	China	20	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Columbia	6	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Cuba	16	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Dominican-Republic	14	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Ecuador	3	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	El-Salvador	20	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	England	14	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	France	2	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Germany	19	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Greece	9	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Guatemala	10	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Haiti	8	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Honduras	2	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Hong	2	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	India	18	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Iran	12	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Ireland	4	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Italy	11	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Jamaica	20	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Japan	13	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Laos	2	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Mexico	107	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Nicaragua	4	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Outlying-US(Guam-USVI-etc)	1	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Peru	5	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Philippines	32	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Poland	10	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000
	Portugal	5	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000
	Puerto-Rico	18	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000
	Scotland	3	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000
	South	13	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000
	Taiwan	7	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000
	Thailand	5	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000
	Trinadad&Tobago	3	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000
	United-States	4701	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000
	Vietnam	11	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
education_group	10th	151	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	11th	186	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	12th	52	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	1st-4th	25	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	5th-6th	48	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	7th-8th	96	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	9th	72	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Assoc-acdm	165	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Assoc-voc	204	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Bachelors	884	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Doctorate	46	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	HS-grad	1706	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Masters	274	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Preschool	7	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Prof-school	97	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Some-college	1162	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
occupation	Adm-clerical	634	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Armed-Forces	1	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Craft-repair	645	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Exec-managerial	680	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Farming-fishing	161	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Handlers-cleaners	222	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Machine-op-inspct	353	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Other-service	591	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Priv-house-serv	26	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Prof-specialty	686	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Protective-serv	123	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Sales	636	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Tech-support	153	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000
	Transport-moving	264	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000
class_of_work	Federal-gov	142	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000
	Local-gov	344	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000
	Private	3826	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000	.000
	Self-emp-inc	190	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.000
	Self-emp-not-inc	439	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000
	State-gov	233	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000
	Without-pay	1	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
marital_status	Divorced	729	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Married-AF-spouse	4	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Married-civ-spouse	2429	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Married-spouse-absent	62	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Never-married	1649	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Separated	166	.000	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Widowed	136	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
relationship	Husband	2140	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Not-in-family	1320	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Other-relative	152	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Own-child	758	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Unmarried	546	.000	.000	.000	.000	1.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Wife	259	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
ethnicity	Amer-Indian-Eskimo	47	1.000	.000	.000	.000	.000																																
	Asian-Pac-Islander	167	.000	1.000	.000	.000	.000																																
	Black	493	.000	.000	1.000	.000	.000																																
	Other	34	.000	.000	.000	1.000	.000																																
	White	4434	.000	.000	.000	.000																																	
sex	Female	1664	1.000																																				
	Male	3511	.000																																				

		Frequency	Parameter coding (1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
birth_country	Cambodia	3	1.000	.000	.000	.000	.000	.000	.000	.000	.000
	Canada	22	.000	1.000	.000	.000	.000	.000	.000	.000	.000
	China	20	.000	.000	1.000	.000	.000	.000	.000	.000	.000
	Columbia	6	.000	.000	.000	1.000	.000	.000	.000	.000	.000
	Cuba	16	.000	.000	.000	.000	1.000	.000	.000	.000	.000
	Dominican-Republic	14	.000	.000	.000	.000	.000	1.000	.000	.000	.000
	Ecuador	3	.000	.000	.000	.000	.000	.000	1.000	.000	.000
	El-Salvador	20	.000	.000	.000	.000	.000	.000	.000	1.000	.000

A Statistical Approach: Binary Logistic Regression

Omnibus Tests of Model Coefficients

Step 12

	Chi-square	df	Sig.
Step	1157342.647	4	.000
Block	503049331.6	91	.000
Model	503049331.6	91	.000

An Omnibus test – used to check that the final model is an improvement over the baseline

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	869913333[a]	.214	.319
2	763120493[b]	.294	.438
3	729656479[b]	.318	.473
4	665306984[b]	.361	.536
5	650267785[b]	.370	.550
6	640055863[b]	.377	.560
7	629672173[b]	.383	.570
8	622212102[b]	.388	.576
9	614856129[b]	.392	.583
10	610166536[b]	.395	.587
11	608044259[b]	.396	.589
12	606886916[b]	.397	.590

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Model summary showing increase in fit at each step in the Forward LR method

A Statistical Approach: Binary Logistic Regression

Classification table showing overall classification accuracy at final step

Classification Table[a]

Step 12

Observed		Predicted					
		Selected Cases[b]			Unselected Cases[c,d]		
		Over_50K_Flag		Percentage Correct	Over_50K_Flag		Percentage Correct
		<=50K	>50K		<=50K	>50K	
Over_50K_Flag	<=50K	698118155	51298386	93.2	693842918	60082401	92.0
	>50K	87695649	157138237	64.2	98880714	139206051	58.5
Overall Percentage				86.0			84.0

a. The cut value is .500

b. Selected cases Approximately 50% of the cases (SAMPLE) EQ 1

c. Unselected cases Approximately 50% of the cases (SAMPLE) NE 1

d. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.

A Statistical Approach: Binary Logistic Regression

Model Coefficients Table

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	relationship			14014784[illegible]	5	.000	
Step 12[1]	age	.026	.000	7232435.244	1	.000	1.027
	education_group			26740709.46	15	.000	
	education_group(1)	-.452	.001	419681.776	1	.000	.637
	education_group(2)	-1.659	.001	2811922.141	1	.000	.190
	education_group(3)	-.811	.001	327524.083	1	.000	.444
	education_group(4)	-1.329	.002	333016.760	1	.000	.265
	education_group(5)	-1.717	.002	1087416.942	1	.000	.180
	education_group(6)	-1.755	.001	2118396.879	1	.000	.173
	education_group(7)	-1.975	.002	1536382.625	1	.000	.139
	[illegible]	[illegible]	[illegible]	[illegible]	1	[illegible]	[illegible]
	capital_gain	.000	.000	32988938.67	1	.000	1.000
	capital_loss	.001	.000	8470975.913	1	.000	1.001
	Constant	-16.342	68.400	.057	1	.811	.000

a. Variable(s) entered on step 1: relationship.

A Statistical Approach: Binary Logistic Regression – note the optional elements that can be added to the analysis

Table showing effect on model if term removed

Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	relationship	-554968124	240022914.9	5	.000
	capital_loss	-308320446	8596632.754	1	.000
Step 12	age	-307092506	7298095.196	1	.000
	education_group	-320845227	34803538.93	15	.000
	hours_per_week	-308360457	9833997.897	1	.000
	class_of_work	-305635982	4385048.049	6	.000
	marital_status	-307448310	8009703.723	6	.000
	occupation	-314036147	21185377.26	13	.000
	relationship	-308304259	9721601.951	5	.000
	ethnicity	-304022129	1157342.647	4	.000
	sex	-304504307	2121699.007	1	.000
	birth_country	-310200538	13514160.89	37	.000
	capital_gain	-336421848	65956779.85	1	.000
	capital_loss	-307791902	8696888.552	1	.000

A Statistical Approach: Binary Logistic Regression – note the optional elements that can be added to the analysis

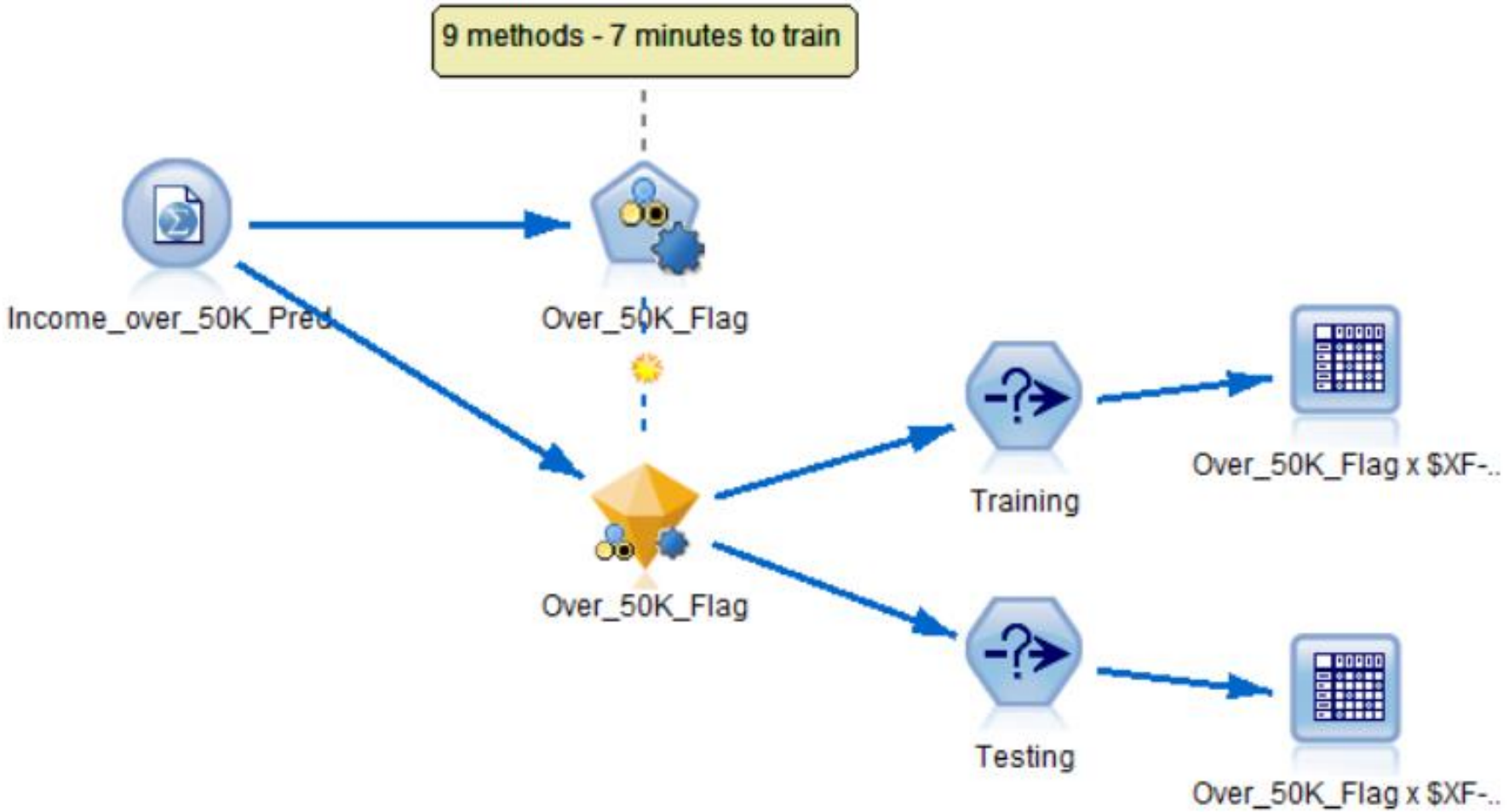
Table showing variable not in the equation at each step

Variables not in the Equation[a]

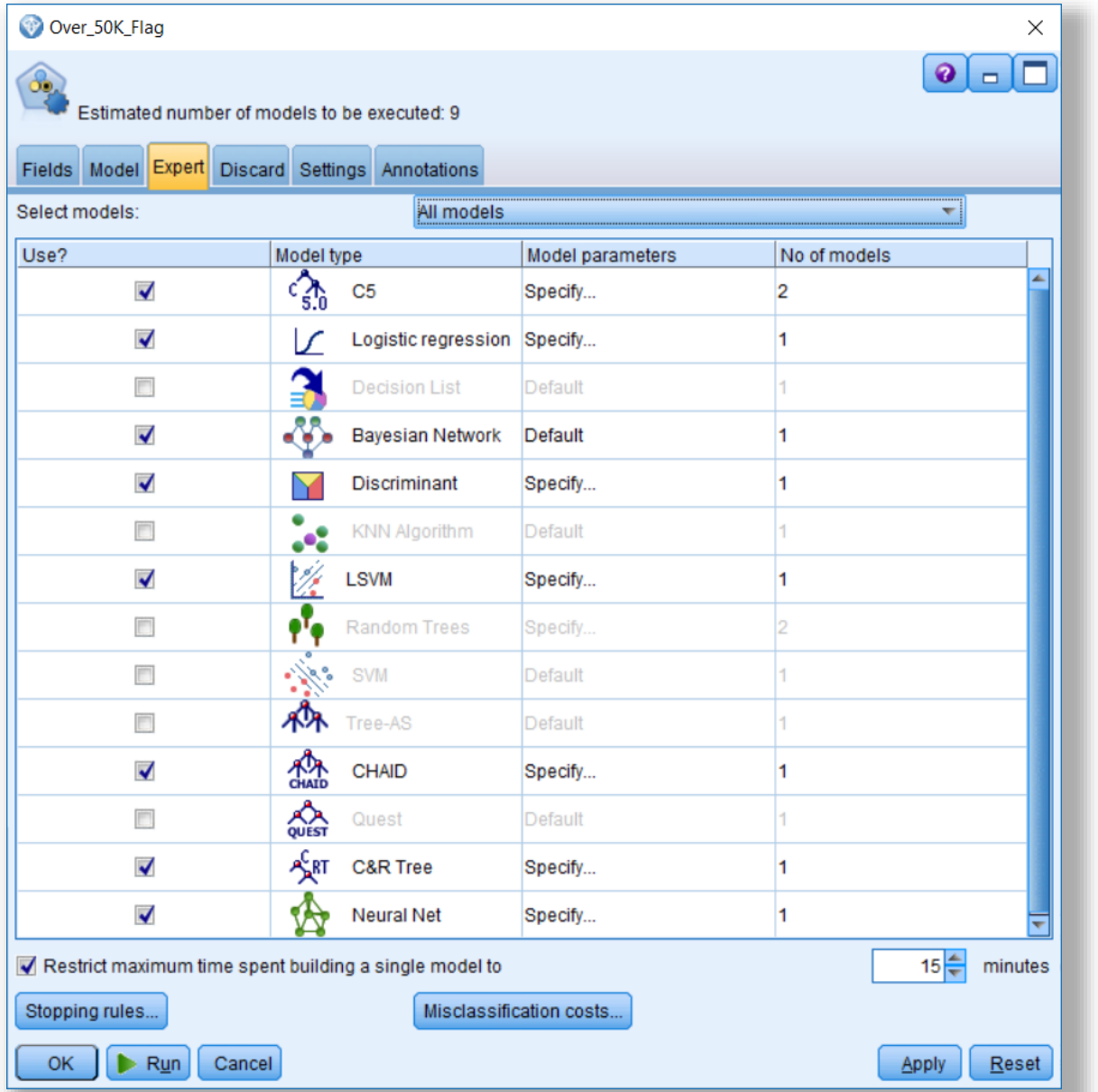
			Score	df	Sig.
Step 1	Variables	age	15580569.05	1	.000
		education_group	101076295.8	15	.000
		education_group(1)	2297554.838	1	.000
		education_group(2)	6448108.856	1	.000
		education_group(3)	1170564.194	1	.000
		education_group(4)	1609203.345	1	.000
		education_group(5)	3141393.367	1	.000
Step 11	Variables	ethnicity	993787.363	4	.000
		ethnicity(1)	964239.685	1	.000
		ethnicity(2)	531.393	1	.000
		ethnicity(3)	13232.836	1	.000
		ethnicity(4)	24104.179	1	.000
	Overall Statistics		993787.363	4	.000

a. Residual Chi-Squares are not computed because of redundancies.

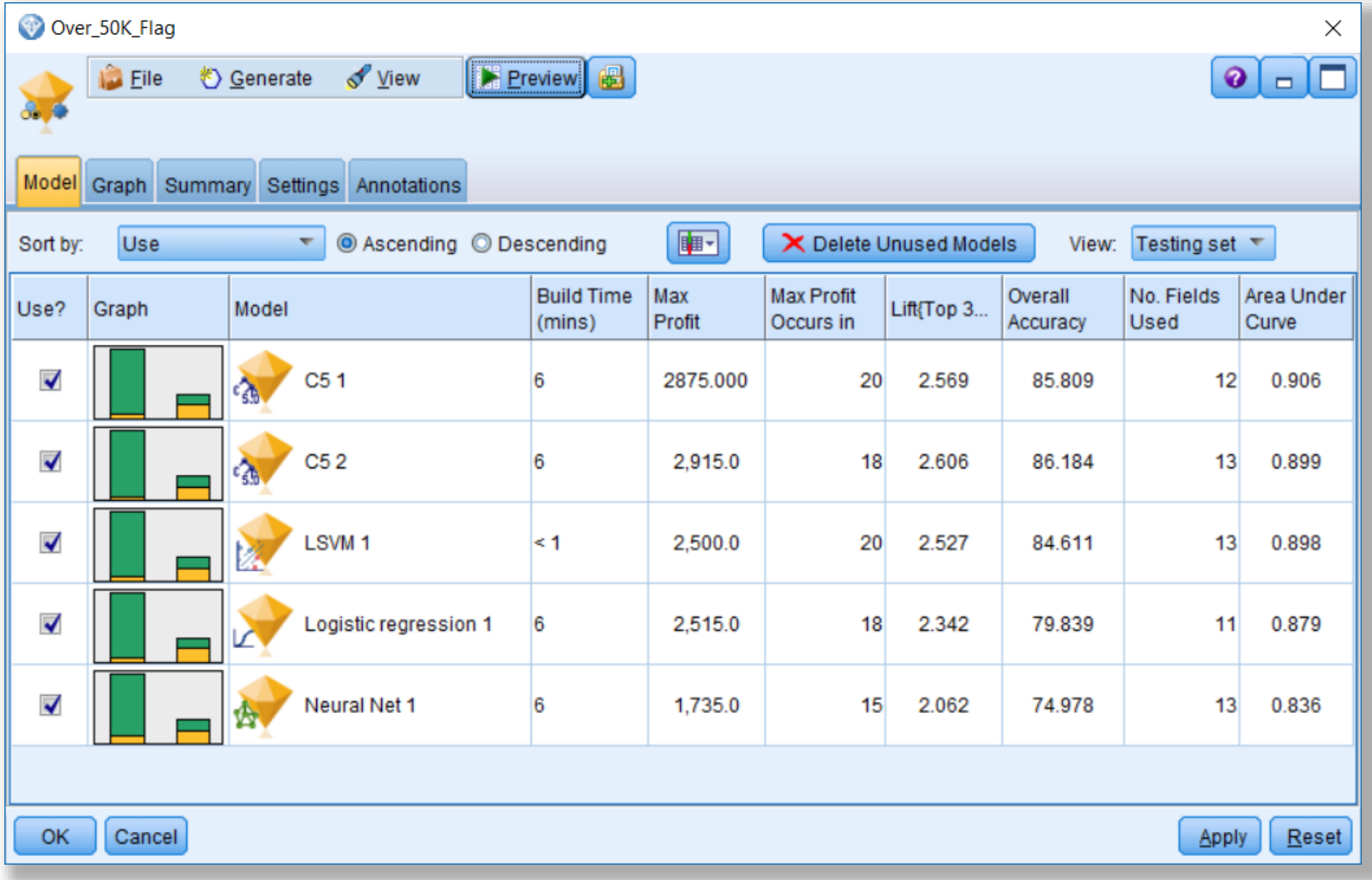
A Data Mining/ML Approach: Multiple methods tested using a automatic classifier



A Data Mining/ML Approach: Multiple methods tested using a automatic classifier

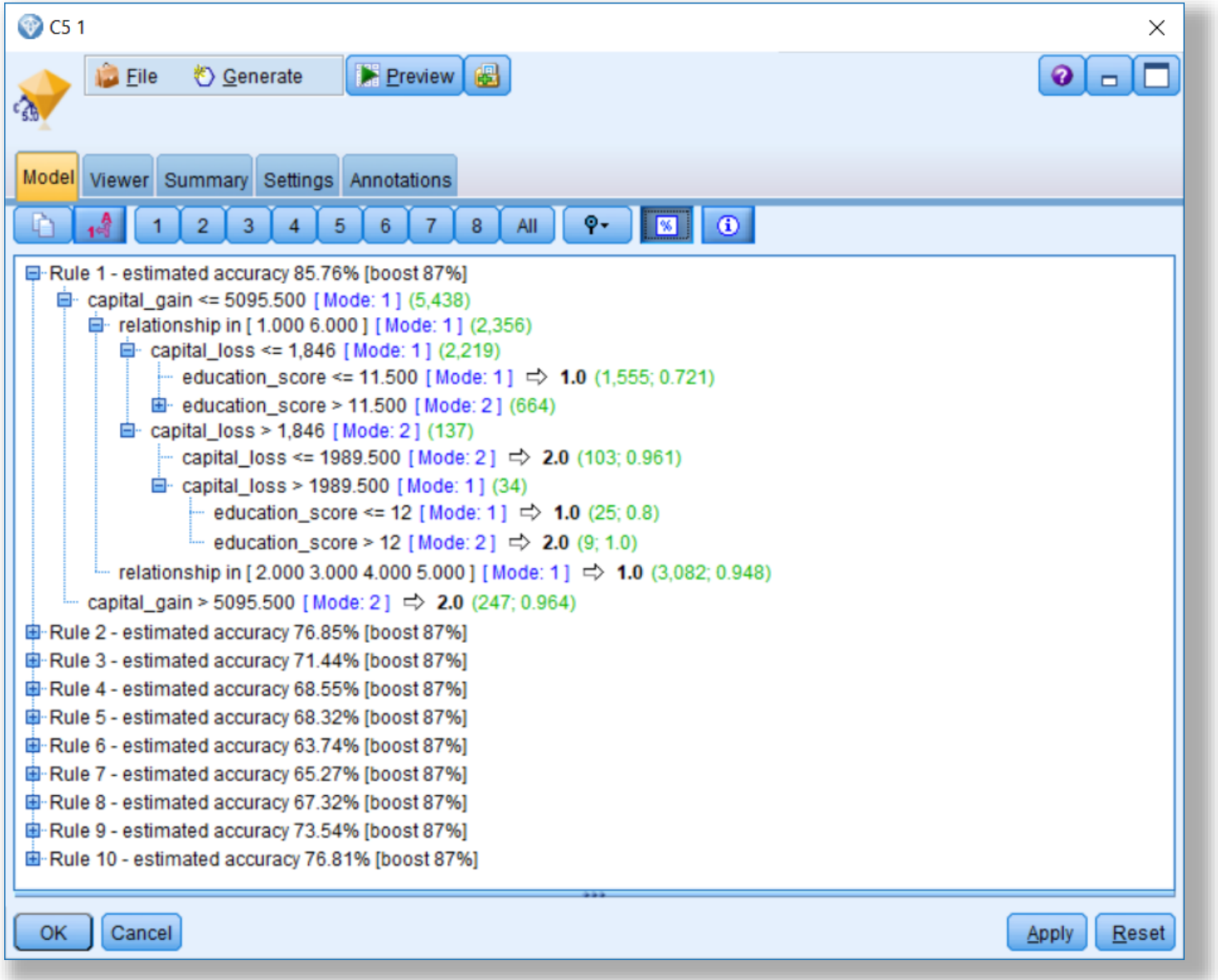


A Data Mining/ML Approach: Multiple methods tested using a automatic classifier



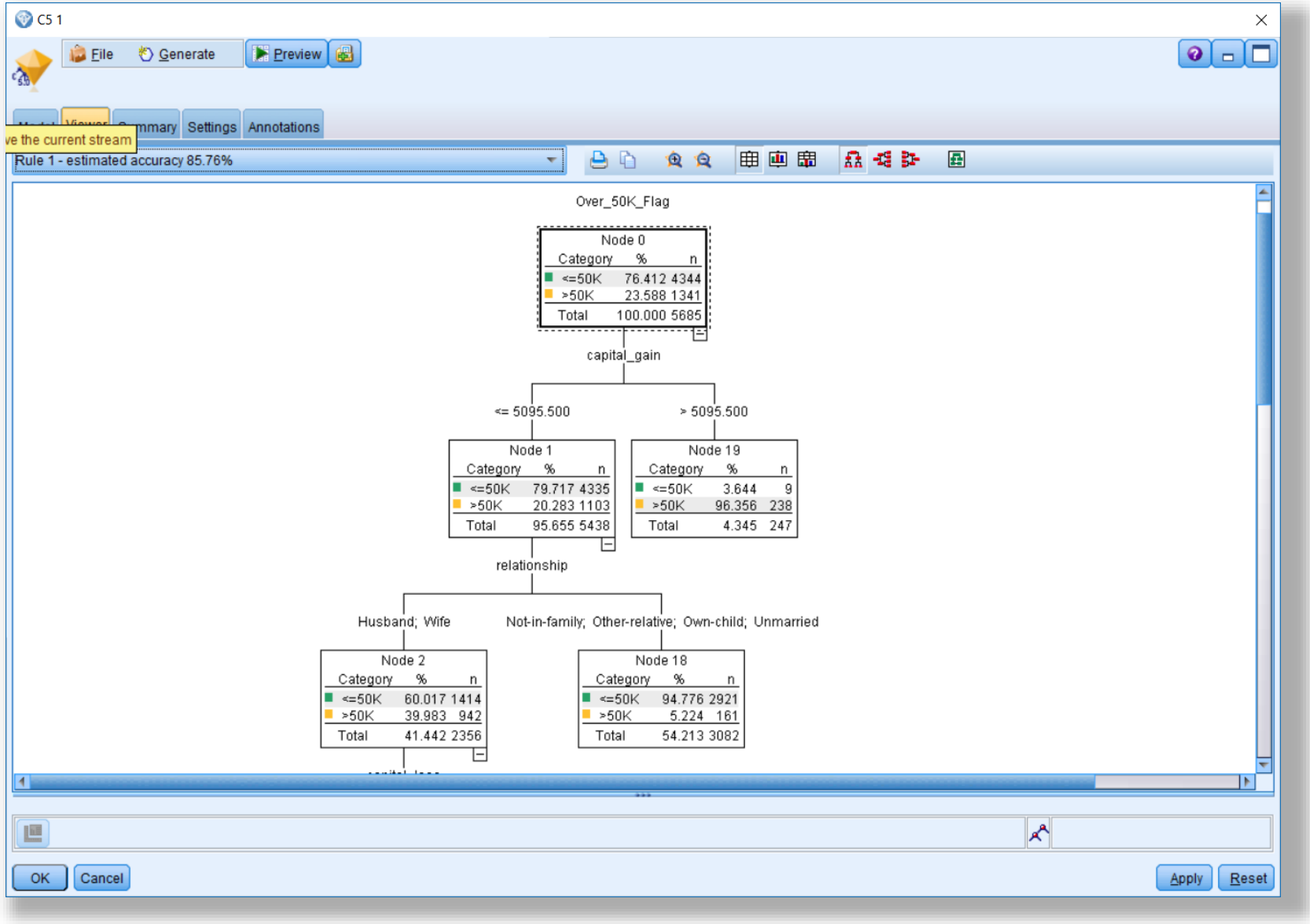
A Data Mining/ML Approach: Multiple methods tested using a automatic classifier

Browsing the first (boosted) C5 model, we can see that it consists of a series of rules that have been generated in 10 separate passes of the data



A Data Mining/ML Approach: Multiple methods tested using a automatic classifier

Each set of rules corresponds to an individual decision tree

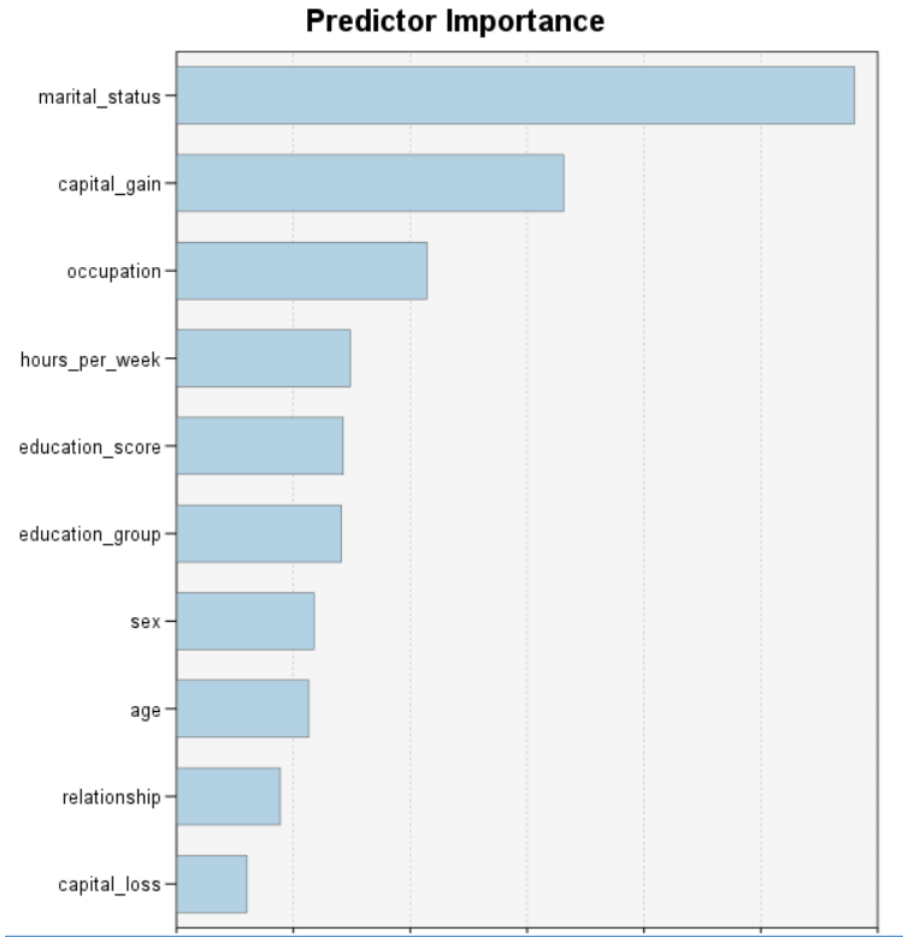


A Data Mining/ML Approach: Multiple methods tested using a automatic classifier

The LSVM (Linear Support Vector Machine) – shows a predictor importance chart and a series of coefficients or "feature weights" similar to a statistical model – but not much else in the way of detailed output

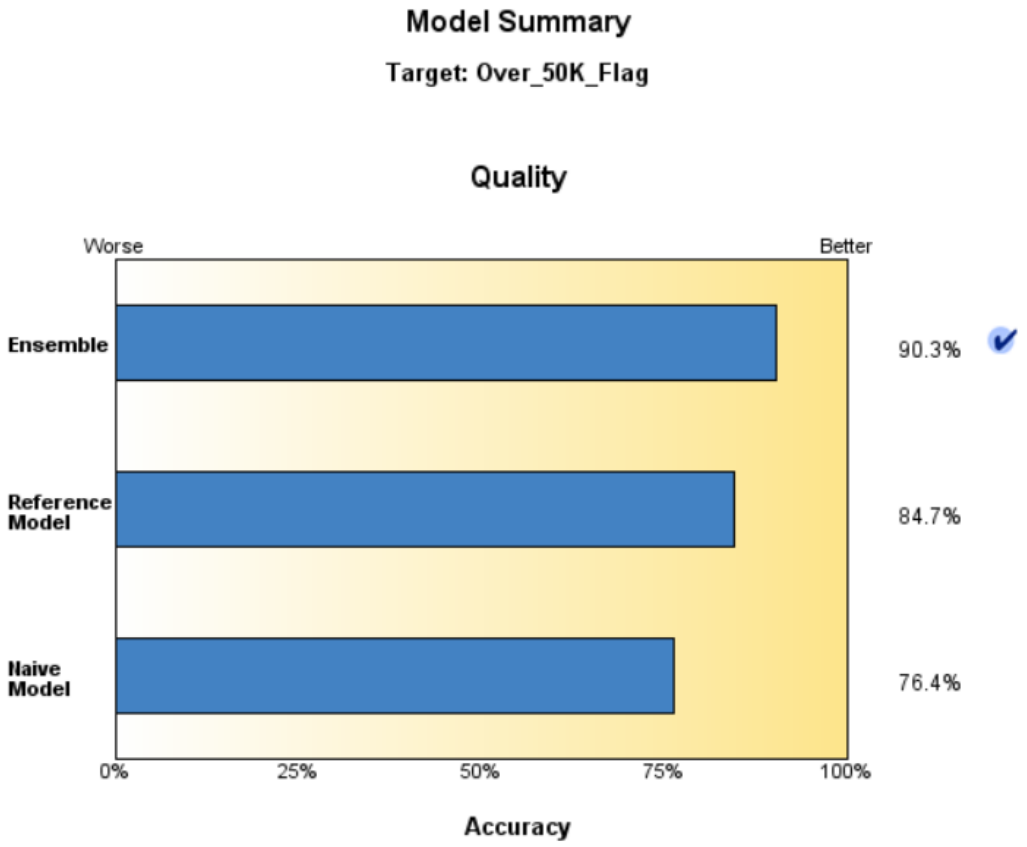
Parameter Estimates

Parameter	B
Intercept	1.2730496178848658
age	-0.009532452768904498
education_score	-0.05143140494774034
hours_per_week	-0.011671566675546786
capital_gain	-0.0001000840007449788
capital_loss	-0.000250083824569036
[education_group=1.0]	0.40259420940334684
[education_group=2.0]	0.3908420626821771
[education_group=3.0]	0.021489391859816822
[education_group=4.0]	0.05811114336436938
[education_group=5.0]	-0.0320470796062458
[education_group=6.0]	0.1118636447657502
[education_group=7.0]	0.3436301228938327
[education_group=8.0]	0.06459725539449182



A Data Mining/ML Approach: Multiple methods tested using a automatic classifier

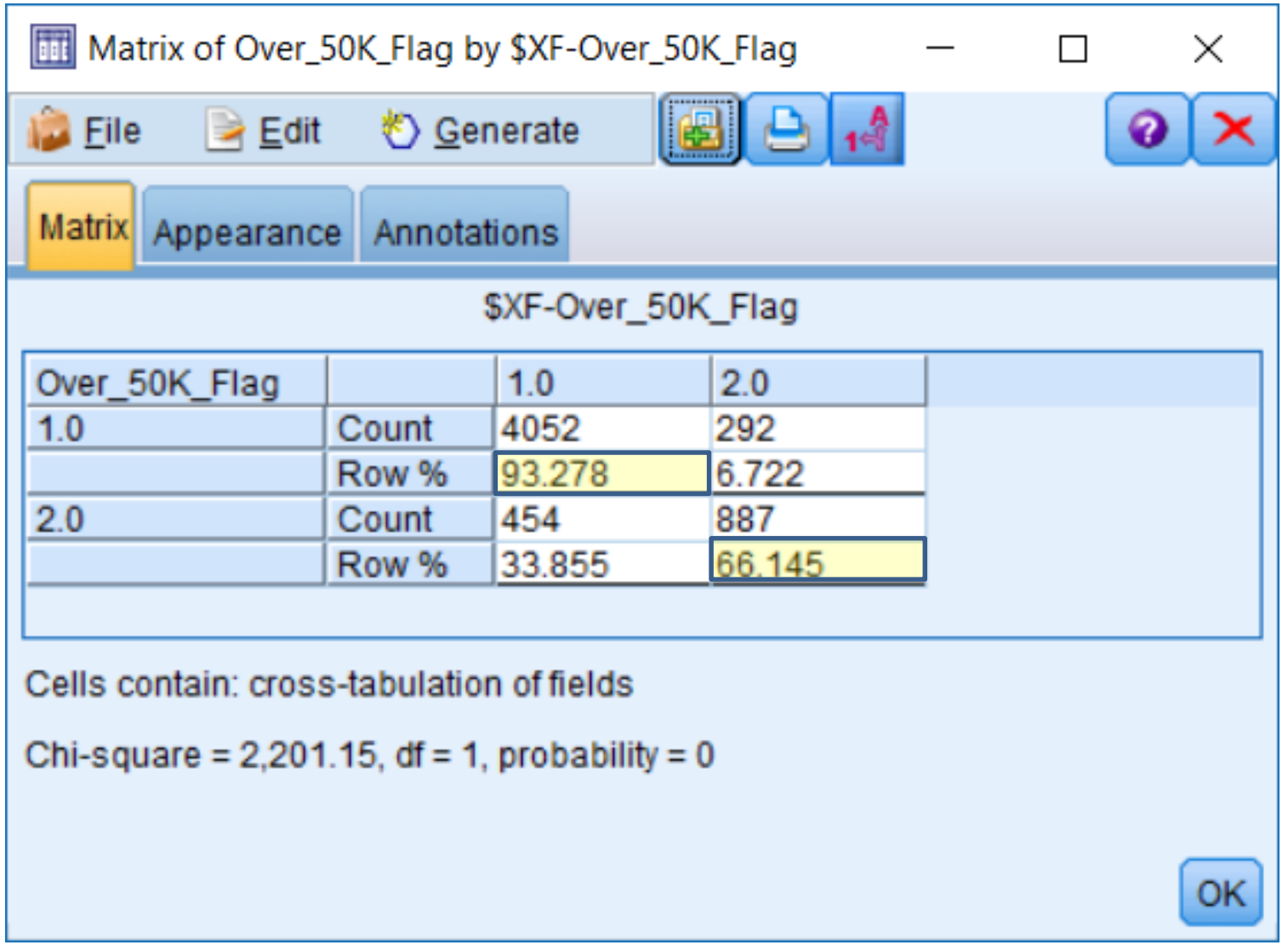
The (bootstrap aggregated) Neural Network shows information about the how often a field was used and the accuracy of the sub-models but *no details* about what those models consist of



Component Model Details

Model	Accuracy	Method	Predictors	Model Size (Synapses)	Records
1	84.7%		13	1135	5,253
2	76.5%		13	674	5,459
3	77.0%		13	1166	5,545
4	73.5%		13	762	5,608
5	75.1%		13	821	5,633
6	66.3%		13	447	5,639
7	70.7%		13	882	5,666
8	72.3%		13	1206	5,669
9	72.6%		13	970	5,671
10	71.7%		13	937	5,680

The C5 boosted models has the highest accuracy on the test group



Did Statistics get 'left behind'?

- "Had we incorporated computing methodology from its inception as a fundamental statistical tool (as opposed to simply a convenient way to apply our existing tools) many of the other data related fields would not have needed to exist. They would have been part of our field." – Jerome H. Friedman from 'Data Mining and Statistics – What's the Connection' -1997

- "The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets." – Leo Breiman from Statistical Modeling: The Two Cultures - 2001

Why Machine Learning matters

Why Machine Learning?

- Certain difficult problems can only be properly addressed with ML
- There are many situations where accuracy is king
- It's often much more flexible than statistical approaches
  - E.g. Neural Networks are not bound by algebraic equations in the same way the Regressions are
- It's where the majority of R&D is focussed
- It is at the heart of hard AI applications
- Because we can. We often have data that is more like the population now. So we don't need samples.

Why Statistics *still* matters

Why Statistics?

- ML is usually a sledgehammer when you need a nutcracker
- There are many situations where transparency is king
- Making inferences about populations from samples is by far the most common analytical problem that people deal with
  - To answer a lot of questions we only have **samples** of data e.g. most ad-hoc surveys, clinical trials and scientific experiments
- Statistics is still 'The Science of Variation' – it's how new things are discovered in data

What can we expect in the future?

Future Predictions

- We see a growing issue in the availability of statistical expertise on the supply side
- ML and AI will continue to grow … and the, arguably purist, distinction between them may disappear
- There will be more ML tools available to Citizen Data Scientists e.g. line of business
  - Cloud-based, often niche and through smarter/simpler User Interfaces
- There will be more productivity and automation
  - There is currently a preponderance of tools that claim to automate data preparation for example