Society of Actuaries in Ireland

Review of the Titanic Competition

15th February 2016

Agenda

- Data Analytics in the Society of Actuaries
- Team ZLAP
- Deloitte GI Team
- Where Can I Get More?

Disclaimer:
The material, content and views in the following presentation are those of the presenter(s).

Data Analytics in the Society of Actuaries

- Wider Fields Committee and Data Analytics subgroup.
- Past events:
  - Who is the driver?
  - Titanic Competition Workshop
- Future events

Disclaimer:
The material, content and views in the following presentation are those of the presenter(s).

Society of Actuaries in Ireland

Titanic: Machine Learning from Disaster
Team ZLAP

15th February 2016

Introduction

- Team ZLAP
  - Nicholas Clarke — Product Solutions
  - Patrick Mangan — Data Analytics
  - Julianne Harrington — Data Analytics

ZURICH

The Problem

- Predict survival on the Titanic

  - Analyse which groups of passengers were likely to survive
  - Apply the tools of machine learning to make predictions about survival
  - Data split into a ‘training set’ and a ‘test set’
  - Training set includes the outcome for each passenger
  - Use training set to build our model to generate predictions for the test set

- IPython Notebook
  - Powerful
  - Fast
  - Flexible
  - Open-source
  - Bundle your analysis in one file
  - A range of packages like **Pandas, NumPy, SciPy, Scikit-Learn, Matplotlib, Statsmodels**

- 891 train / 418 test

- Variables:
  - *Name*
  - *Sex*
  - *Age*
  - *Number of Siblings/Spouses Aboard*
  - *Number of Parents/Children Aboard*
  - *Ticket Number*
  - *Passenger Fare*
  - *Cabin*
  - *Port of Embarkation*

Feature Engineering

- Extracting title from name
- Family grouping
  - Survival status of family members (spouse, parent/child)
- Normalising data
  - Log(fare)
  - Log(fare) outside 2 standard deviations
- Categorical Variables
  - Child
  - Lone traveller

Imputing Missing Variables

- Averaging across sub groups
- Randomised Lasso Regression
  - Modelling ages
  - Automatic feature selection

- Men, women, and children were modelled separately.
  - Allowed for group-specific covariates to be created.
  - Less data in each group for cross-validation.
  - Some covariates have different meanings/strengths for each of the groups.

- Avenues not explored:
  - Ethnicity/language
  - Matching by tickets

Models Used

- Logistic Regression
  - Widely used, reasonably simple classifier.
  - Models the *probability* that a passenger survives.

- Decision Trees
  - Uses consecutive “splitting” rules to classify data points.
  - Tree is then “pruned” (via cross-validation) to avoid over-fitting.
  - Even still, decision trees suffer from high variance!

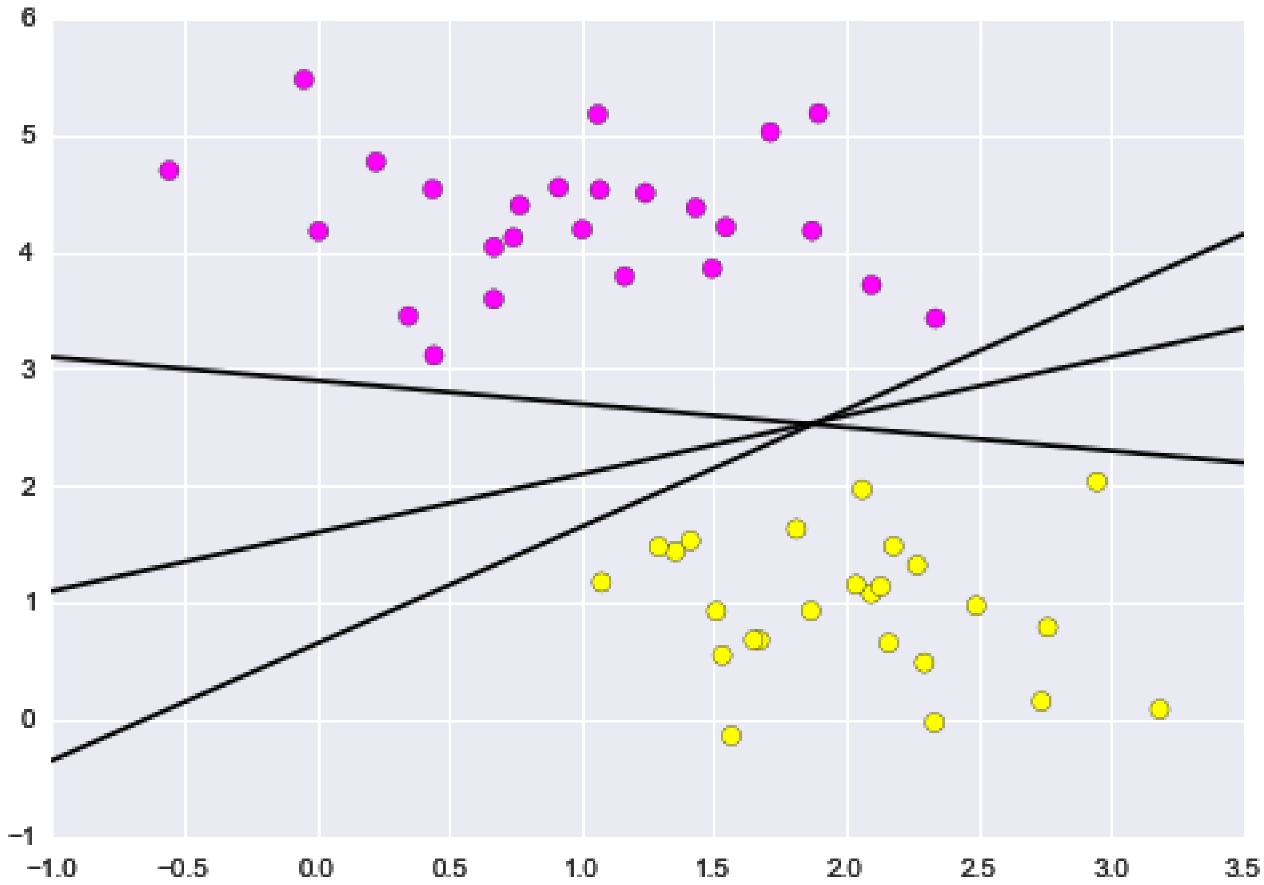
- Bagging / Random Forests
  - Bootstrapping (“bagging”) helps reduce variance.
  - Random Forests then decorrelates the trees.

- Ensemble Learning

Support Vector Machines

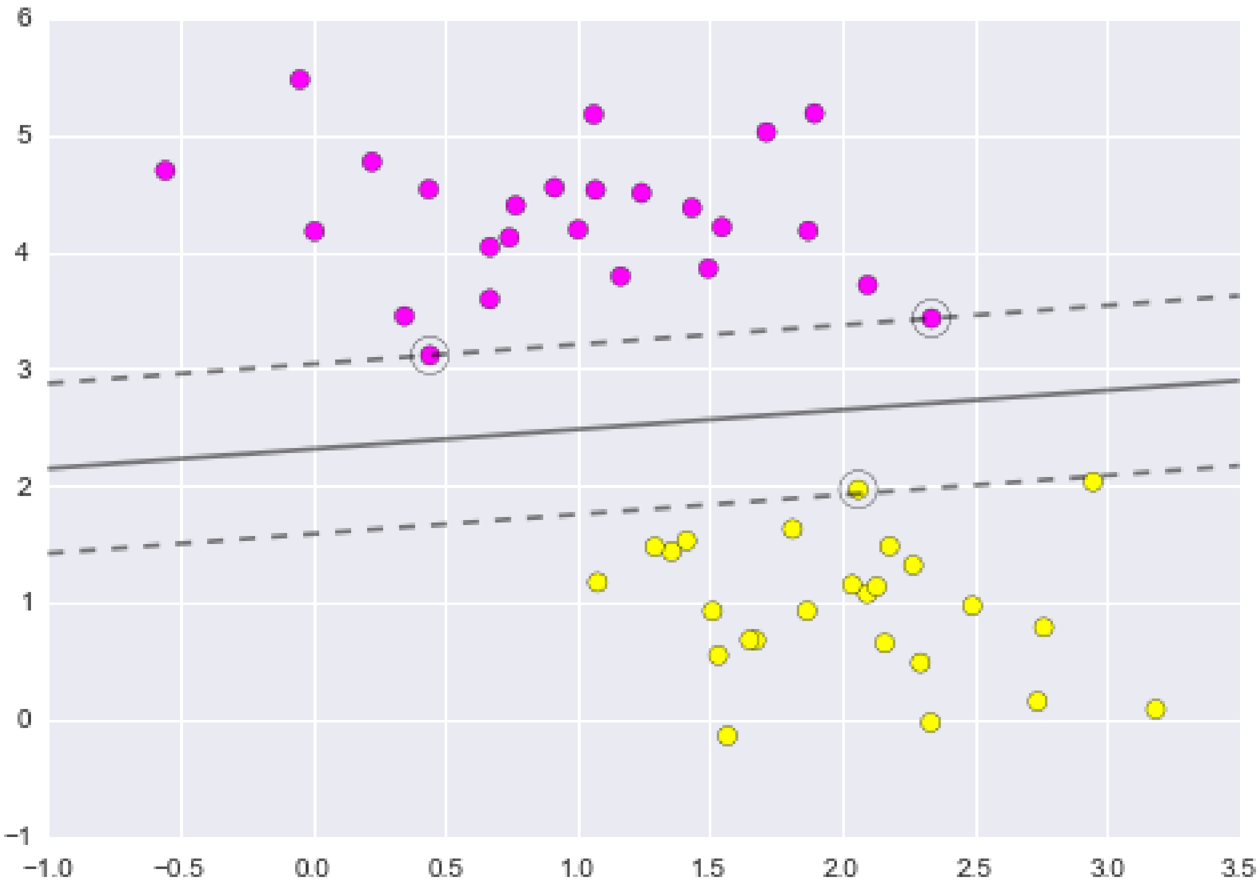
A Simple Classification Problem

We want to find the separating hyperplane.



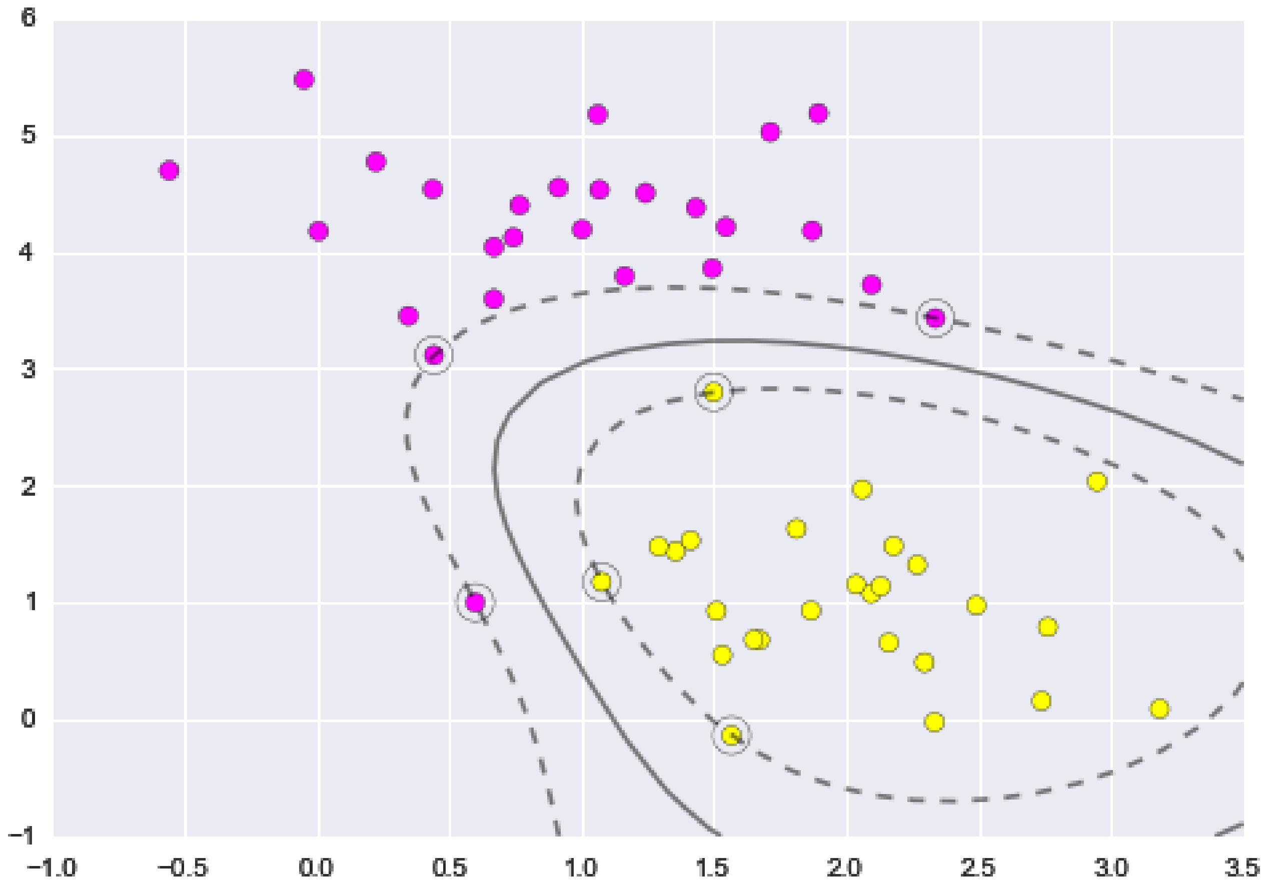
Support Vector Machines

SVM looks for the *maximal margin hyperplane*.



A Slightly Less Simple Classification Problem

Don’t need to restrict to a linear separator.



Support Vector Machines & The Titanic

- Not possible/prudent to correctly classify all training points
  - Some data points will be on the wrong side of the hyperplane.

- How much do we want to avoid misclassification?
  - If 9/10 1^{st} class women survived in our training set, should we predict all 1^{st} class women to survive?

- How much influence should each individual training point have?
  - Does the fate of a 1^{st} class 20 year-old tell us anything about the fate of a 1^{st} class 21 year-old? What about a 30 year-old?

Model Specifics

- Men, women, and children were modelled separately.

- Features used were:
  - Women: *Social class, age, log(fare), log(fare) outside 2sd, title, lone traveller, pensioner, husband's fate, husband's title, children's fate*
  - Children: *Social class, gender, log(fare), log(fare) outside 2sd, age, toddler, mother's fate, father's fate, father's title, siblings' fate, lone traveller*
  - *Men: Gender…*

Result

- Our Score

  – Public Score: 0.82297

    i.e. our model correctly predicts survival for 82.3% of the passengers

	A	B	C	D	E
1	PassengerID	Survived			
2	892	0			
3	893	0			
4	894	0			
5	895	0			
6	896	1			
7	897	0			
8	898	0			
9	899	0			
10	900	1			
11	901	0			
12	902	0			
13	903	0			
14	904	1			
15	905	0			
16	906	1			
17	907	1			
18	908	0			
19	909	0			
20	910	0			
21	911	0			
22	912	0			
23	913	1			

testres

Ready

Any Questions?

Society of Actuaries in Ireland

Titanic: Machine Learning from Disaster
Deloitte GI

15th February 2016

Deloitte Model - Introduction

- Team introduction
- Overview of software used
- Overview of general approach
- Challenges
- Next steps / future improvements

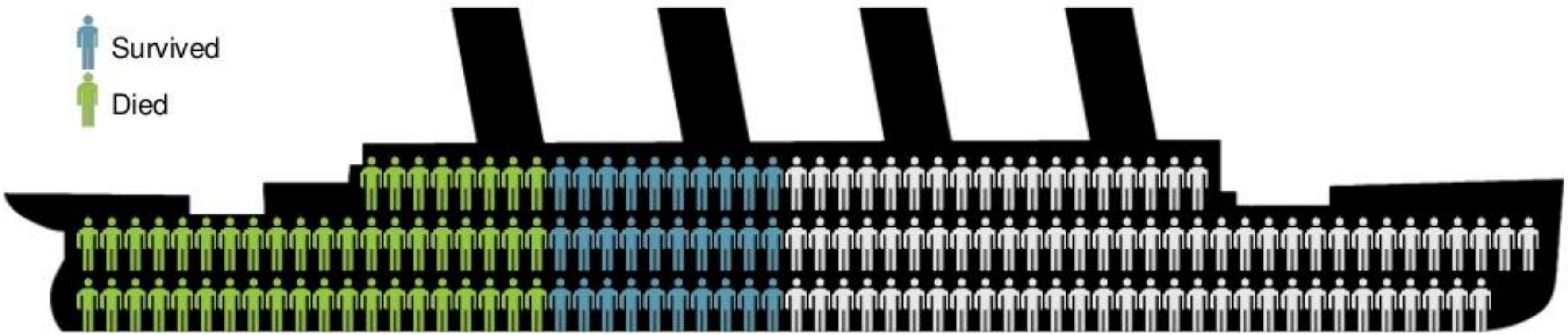
Software and Resources

- Excel
  - Exploratory analysis
  - One-way tables, two-way tables

- R
  - Feature engineering
  - Data adjustments
  - Model training
  - Model testing
  - Model output for submission to Kaggle

- Useful Resources
  - Kaggle tutorial and forums
  - R help files
  - SAI workshop

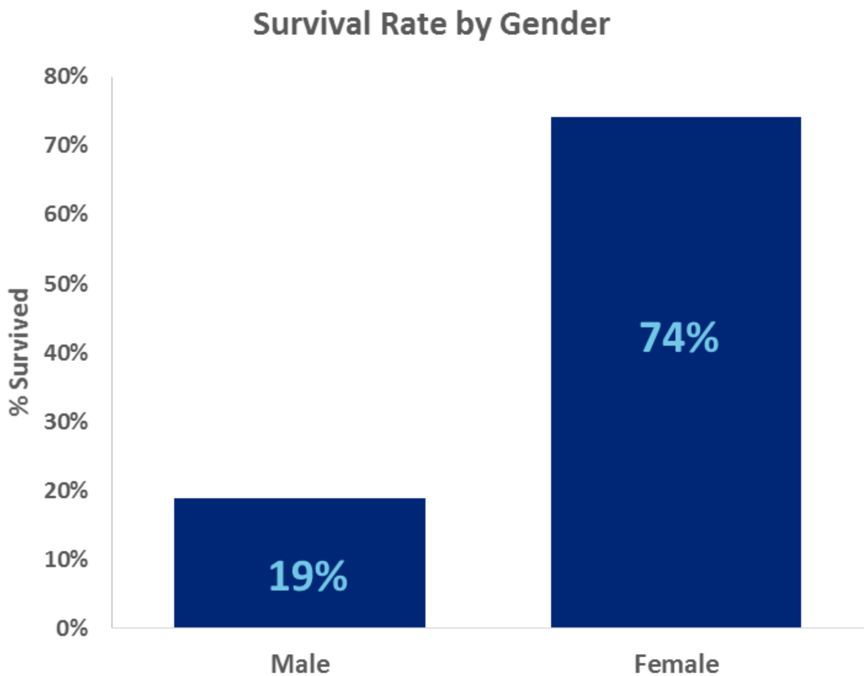
Exploratory analysis



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
31	0	1	Uruchurtu, Don. Manuel E	male	40	0	0	PC 17601	27.7208		C
246	0	1	Minahan, Dr. William Edward	male	44	2	0	19928	90	C78	Q
746	0	1	Crosby, Capt. Edward Gifford	male	70	1	1	WE/P 5735	71	B22	S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S

Exploratory analysis

- One-way and two-way tables used to identify variables of statistical significance



	Class		
	1	2	3
Count	216	184	491
# Survived	136	87	119
% Survived	63.0%	47.3%	24.2%

	Age Group		
	Adult	Child	Missing
Count	601	113	177
# Survived	229	61	52
% Survived	38.1%	54.0%	29.4%

- Missing and incomplete data fields were identified e.g. Age, location embarked, fare.

Feature engineering

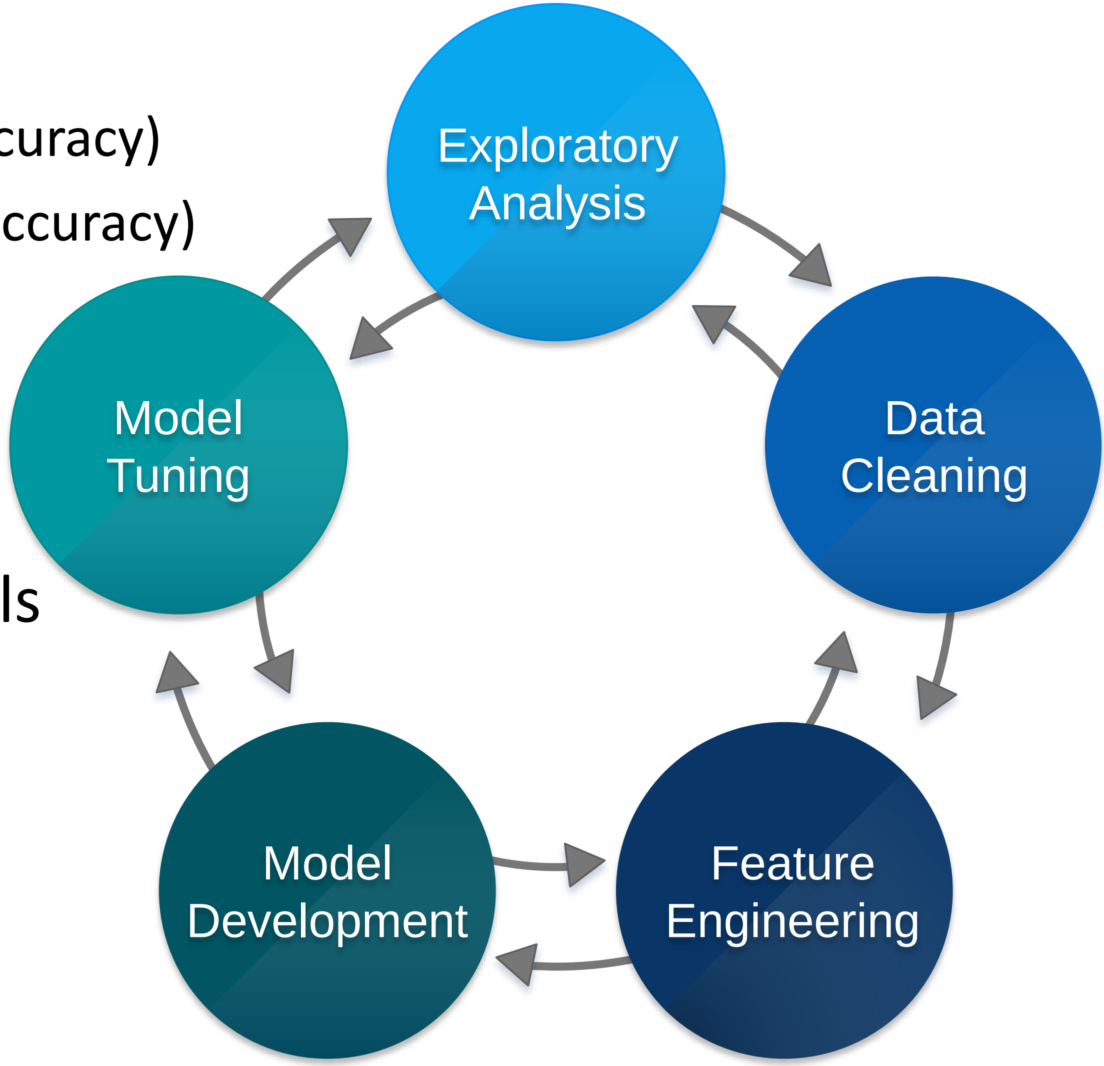
- Engineered new variables based on the data available:
  - **Title:** Indicator of sex and age.
    - Extracted from passenger name
    - Less common/rare titles grouped *e.g. 'Capt', 'Don', 'Major' grouped in with 'Sir'.*
  - **Family Size:**
    - # of siblings + # of parents + 1
  - **Family ID:**
    - Family name & size
    - "Small" for 2 or less (or erroneous data)

Data adjustments

- Data adjustments were carried out in R, to estimate missing and incomplete data items:
  - **Age**:
    - ~20% of passengers have blank ages
    - Filled in blanks using decision tree (utilised engineering variables)
    - *Key data adjustment.*
  - **Location Embarked:**
    - Information for two passengers missing –assumed embarked at most popular location (Southampton).
  - **Fare:**
    - One fare missing – assumed median fare.

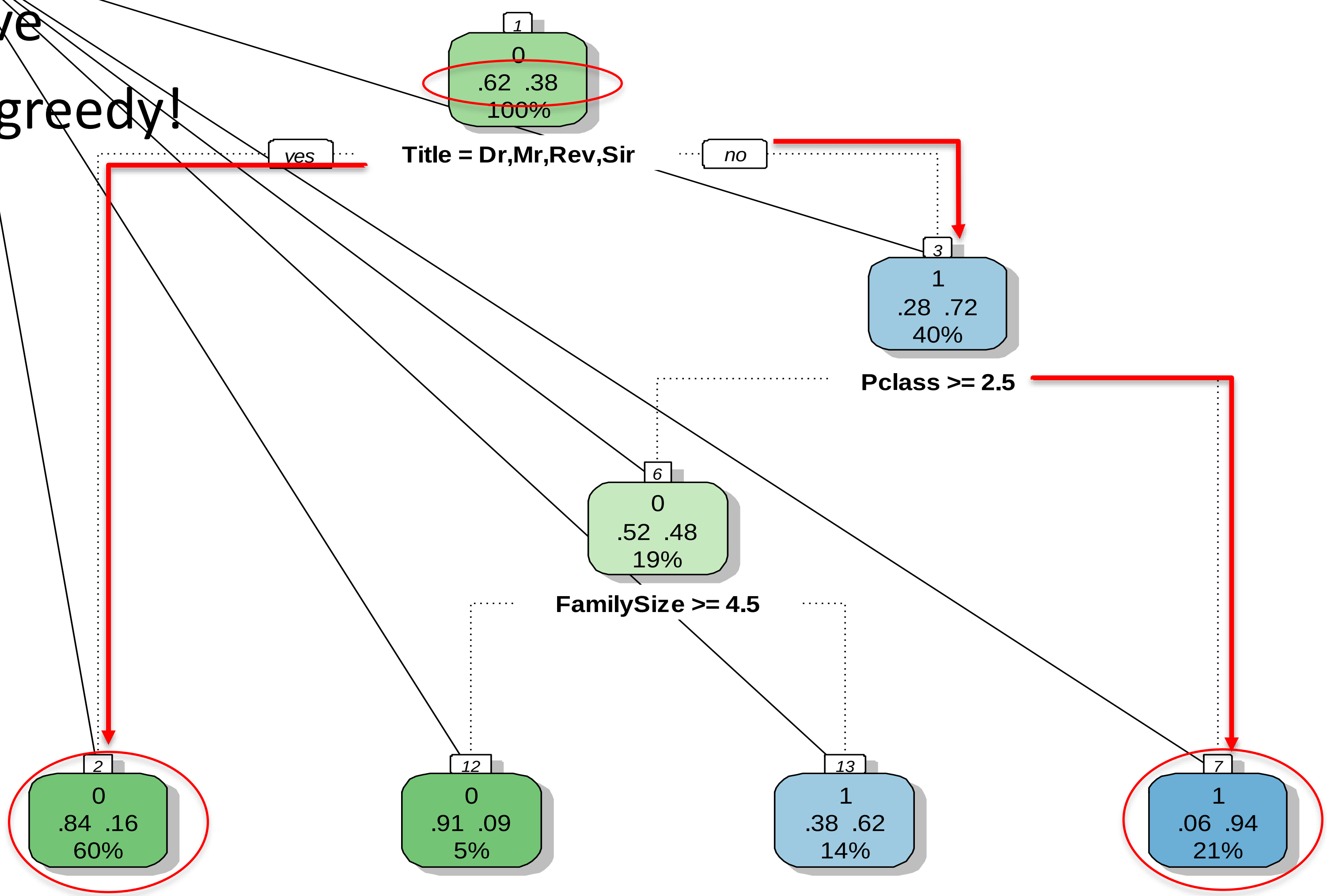
Model training – An iterative process!

- Early models
  - Everyone dies! (~62% accuracy)
  - Women survive (~ 74% accuracy)
- Machine learning models
  - Decision tree
  - Binomial regression
  - Random forest



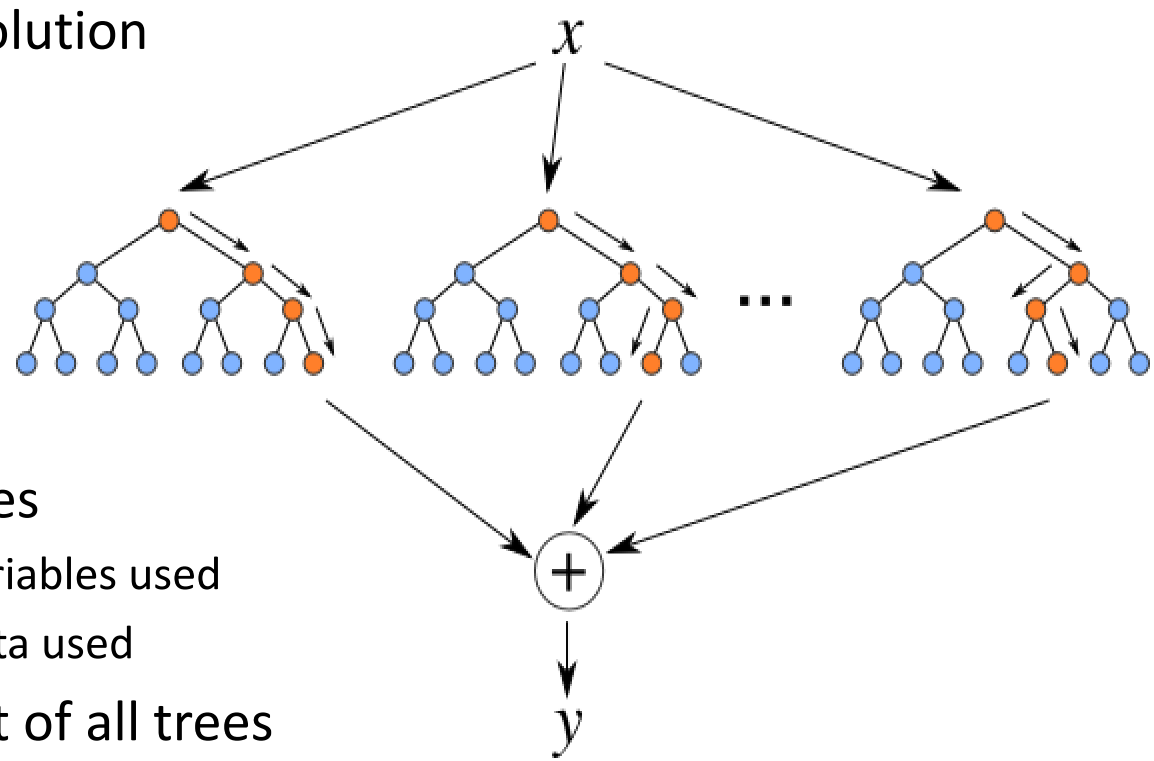
Model training – Decision tree

- Set of rules
- Intuitive
- <u>BUT</u>… greedy!



Model training – Random Forests

- Problem with decision trees
  - May miss 'optimal' solution
  - Prone to overfitting



- Random forests
  - Multiple decision trees
    - Random subset of variables used
    - Random subset of data used
  - Returns **mode** output of all trees
  - Corrects for overfitting

Model training – Binomial Regression

- $\eta_i = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Title}_i + \cdots + \beta_5(\text{Sex}_i * \text{Class}_i) + \cdots$

- $\mathbb{P}(\text{Survived}) = \frac{1}{1+e^{-\eta_i}}$

- Predict passenger survived if $\mathbb{P}(\text{Survived}) > 0.55$

- 0.55 threshold based on value which maximised

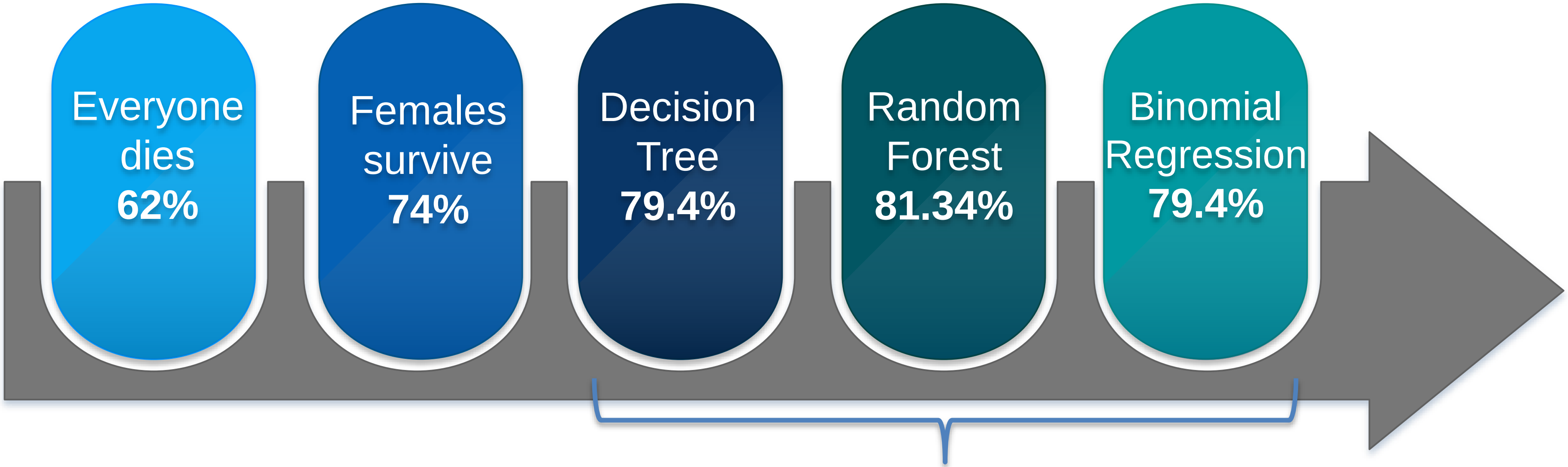
$$\text{Accuracy} = \frac{(\#\text{ of True }-\text{ve}) + (\#\text{ of True }+\text{ve})}{\text{Total }\#\text{ of observations}}$$

Model training – Combining models

- Final model = vote across the 3 models
  - 0/3 or 1/3 survive –> DIED
  - 2/3 or 3/3 survive –> SURVIVE

Passenger ID	Model 1 DT	Model 2 RF	Model 3 GLM	Model 4 Combined
893	✓	✓	✓	✓
894	✗	✓	✗	✗
895	✗	✓	✓	✓
896	✗	✗	✗	✗
...	...	...	...	...

Set of weak learners = strong learner?



Combined Model Result: 81.8%

Possible Next Steps

- Limitations existed:
  - Time
  - Resource
- Possible next steps / enhancements:
  - Further cleansing of data
  - Enhanced feature engineering
  - Further model testing, identifying insignificant variables.
  - Combining algorithms
  - Additional algorithms – e.g. LDA

Conclusion

- Key step: data cleaning, feature engineering
- Diminishing marginal returns of predictive power
- Furthered knowledge of machine learning and R
- Actuarial skillset highly transferable to data analytics

Where Can I Get More?

- Formal education: UCD Msc Data Analytics, UCD Business School MSc in Business Analytics, DIT Msc Computing (Data Analytics)

- Web: Kaggle, KDNuggets, UCI Machine Learning Repository, R-Bloggers, numerous sites for online courses such as Coursera, LinkedIn groups, etc.

- MeetUp Groups: Dublin R, Data Scientists Ireland, Deep Learning Dublin, Dublin Data Science Beginners, Machine Learning Dublin, Hadoop User Group Ireland, and many more!

- Dublin R: San Francisco Crime Database exploration 24th February.

Disclaimer:
The material, content and views in the following presentation are those of the presenter(s).