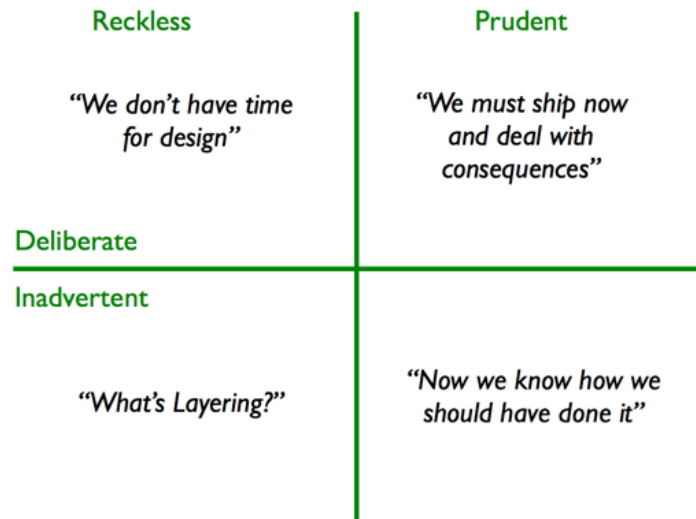


Machine Learning

High Interest Credit Card of Technical Debt.



Software Engineering, dealing with Technical Debt.



TechnicalDebtQuadrant, Martin Fowler

<http://martinfowler.com/bliki/TechnicalDebtQuadrant.html>

I ♥
refactoring

- 10.) Lousy Comments
- 9.) Trailing Whitespace
- 8.) Commented-Out Code
- 7.) Needless Parentheses
- 6.) Powerless Code
- 5.) Unnecessary Requires
- 4.) The Booleaneast Boolean
- 3.) Too Much Hard Work
- 2.) Duplicated Tests
- 1.) Combine All the Codejunk

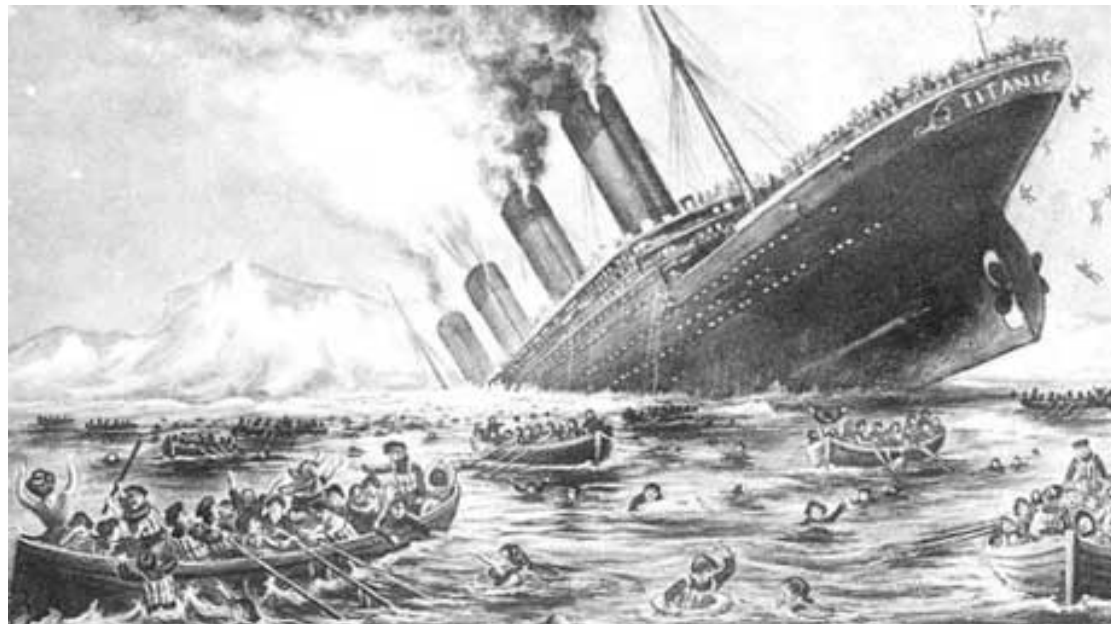
Therapeutic, Katrina Owen

<http://puppetlabs.com/top-20-flowcon-slides-continuous-delivery>

- Strong abstraction boundaries help create maintainable code in which it is easy to make **isolated changes**.
- Machine Learning packages have all the basic code complexity issues as normal code. **But at a system-level, a machine learning model may subtly *erode abstraction boundaries*.**

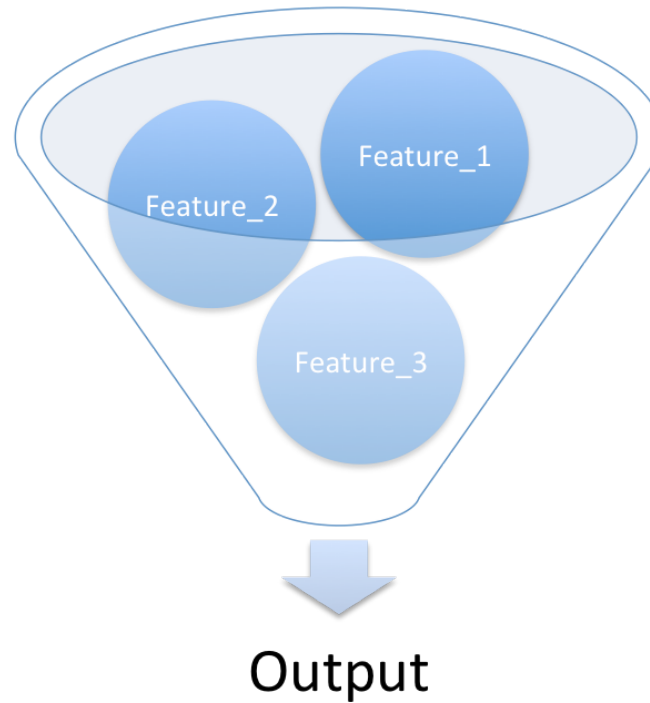
Entanglement

Changing Anything Changes Everything



Postcard showing the Titanic sinking, Rex Features

- Our running example will be using Titanic dataset available on Kaggle website. The objective for this sample competition is to predict survivors. However, we will extend this task by asking other similar questions. The objective is to demonstrate some of the bad practices in building ML models.



A typical Machine Learning Project!

Out[522]:

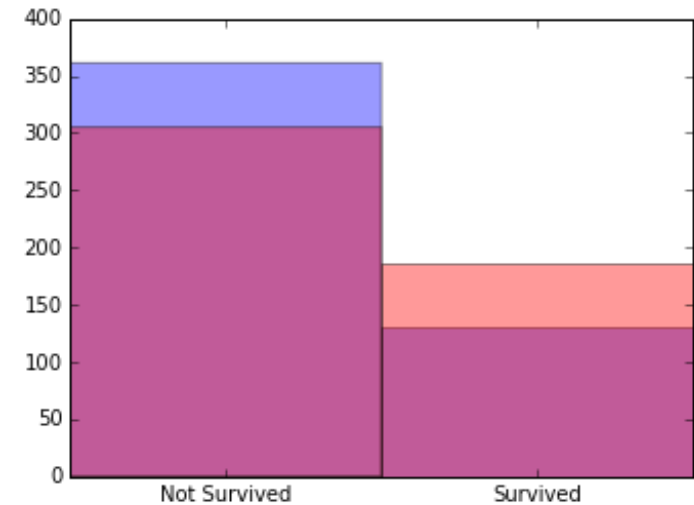
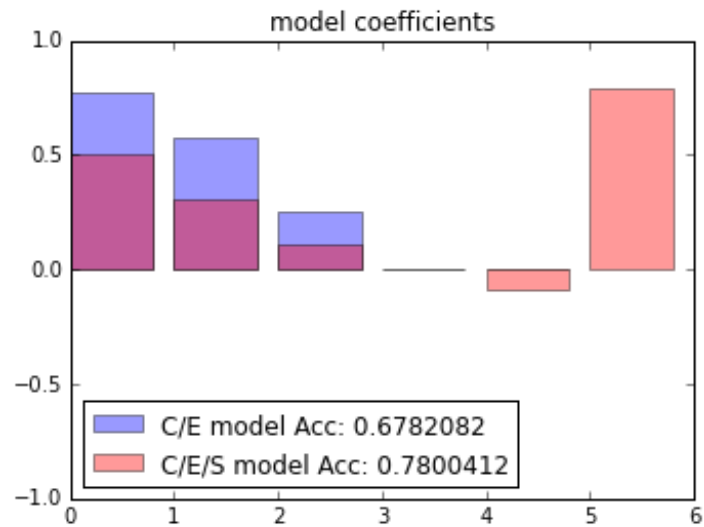
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th	female	38	1	0	PC 17599

										
					Heikkinen,						
2	3		1	3	Miss.	female	26	0	0	STON/O2.	
					Laina					3101282	

3 rows × 12 columns

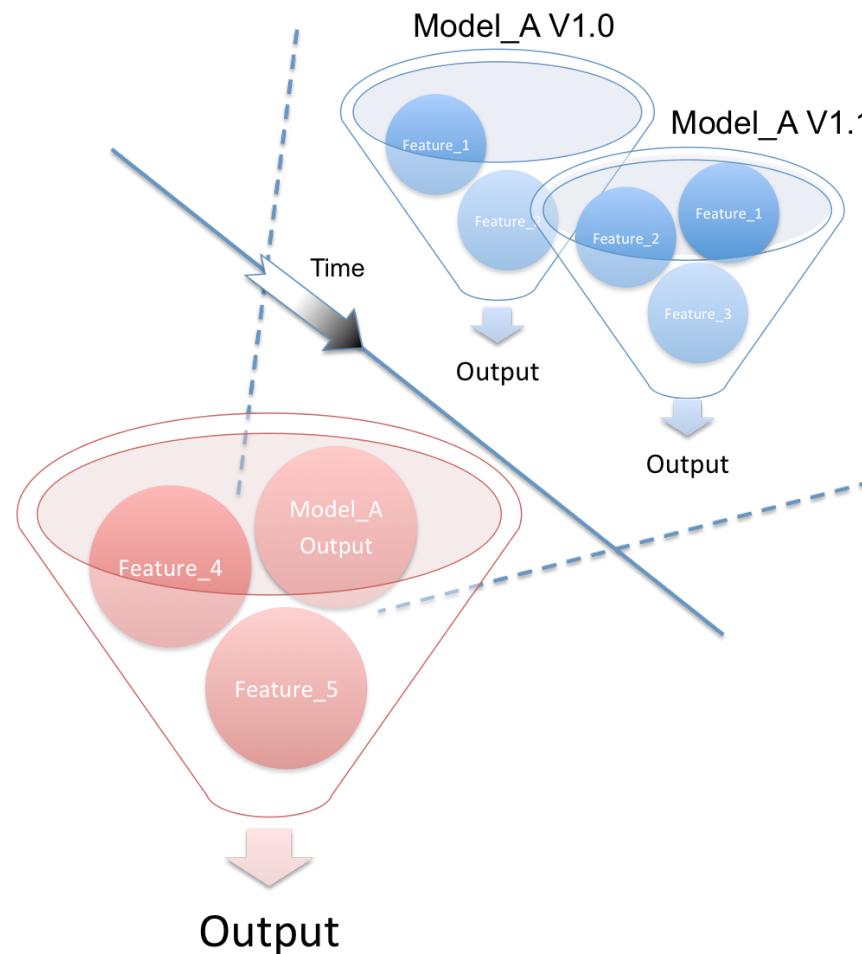
Who Survives?

Out[515]:



Undeclared Data Dependency

Predict age of the passengers

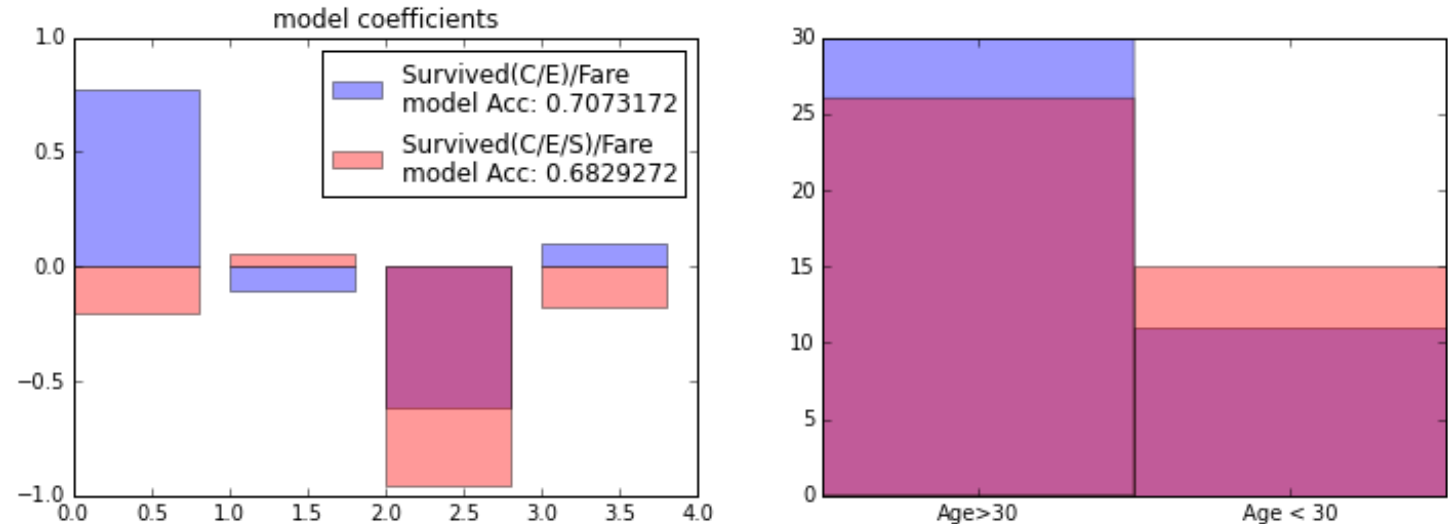


Undeclared customer of a Machine Learning Project!

- Our next question is to predict the age of passengers. A simplified version of this question is to classify passengers older than 30 and younger.
- Here we use the prediction of the previous step. One may assume knowing if someone has survived or not can help determining the age of the person.
- We will be using two kinds of signals. We use the output of the model based on *Passenger Class* and *Port of Embarkment* to train our classifier.
- At the next step we use the output of **modified** model based on Passenger Class and Port of Embarkment and Sex to make our predictions.
- Since the underlying mechanism has changed, our result will be worse. Retraining the

- Since the underlying mechanism has changed our result will be worse. Retraining the model using modified model will improve the result. We can also observe the change in coefficients as a consequence of change in underlying input data.

Out[523]:



Model	Training data	Test data	Acc
Survived(C/E)/Fare	Survived(C/E)/Fare	Survived(C/E)/Fare	0.7073
Survived(C/E)/Fare	Survived(C/E)/Fare	Survived(C/E/S)/Fare	0.5122
Survived(C/E/S)/Fare	Survived(C/E/S)/Fare	Survived(C/E/S)/Fare	0.6829