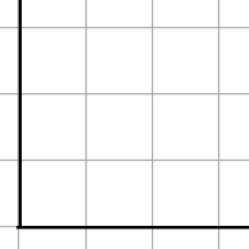
Titanic Survival

### Today, we’re going to create a model to predict who would have lived (and died) during the sinking of the titanic. Find a partner, and you will pair up together to build the best predictor. The top three teams will receive some AWESOME Prizes!

# Before we Start

1. List your dependent variable:
2. List a few independent variables:
3. What type of question are we asking?
4. What type of algorithm will we want to use?
5. What does a good model look like when charted?

# Get the data

1. Navigate to <http://www.kaggle.com> and Log in
   1. Use the credentials you set up for homework last night.
2. Scroll down to the Titanic Example, and click on it

Machine generated alternative text:
ting 
•ted 
Titanic: Machine Learning from Disaster 
Predict survival on the Titanic using Excel, Python, R & Random Forests 
Knowledge 
2 months 
3372 teams 
1475 scripts 
Knowledge 

1. Click the data tab on the right side

Machine generated alternative text:
Dashboard 
Home 
Data 
Make a submission 
Information 
Knowledge • 3,372 teams 
Titanic: Machine Learning from Disaster 
Fri 28 sep 2012 
Thu 31 Dec 2015 (2 months to go) 
Competition Details » Get the Data » Make a submission 
Predict survival on the Titanic using Excel, 
Python, R & Random Forests 

1. Download the train CSV file, save it, then open it in Excel

Machine generated alternative text:
Dashboard 
Home 
Data 
Make a submission 
Information 
Description 
Evaluation 
Rules 
Prizes 
Frequently Asked Questions 
Further Reading / Watching 
Knowledge • 3,372 teams 
Titanic: Machine Learning from Disaster 
Fri 28 sep 2012 
Competition Details » 
Data Files 
File Name 
train 
gendermodel 
genderclassmodel 
Get the Data 
Thu 31 Dec 2015 (2 months to go) 
Make a submission 
Available Formats 
:.csv (59.76 kb)t 
.csv (3.18 kb) 
.csv (3.18 kb) 



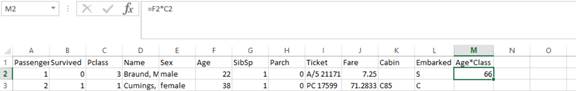
# Feature Engineering

 Add a column for another feature.

* 1. Let's try **Age\*Class**, and put that text in M1

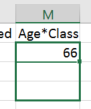
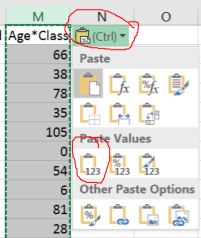


* 1. Enter the following formula into cell M2 =F2\*C2



* + 1. This multiplies the first Feature Vector of Age by the Feature vector of Class
    2. Why would we do this?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* 1. Grab the bottom right corner of that cell, and drag it all the way down to end of the data (row 892).
     1. This will populate the formula down the entire length of the column, and do all the calculating for us.
  2. We now need to keep the data but clear the formula. This way, our Algorithm will be able to understand the column.
     1. Select all of column M, and copy it.
     2. We are going to paste them on the same area we just copied by pressing **cmd-V**. A small box will pop up in the bottom right corner you'll have to click it and select paste values

## PAUSE

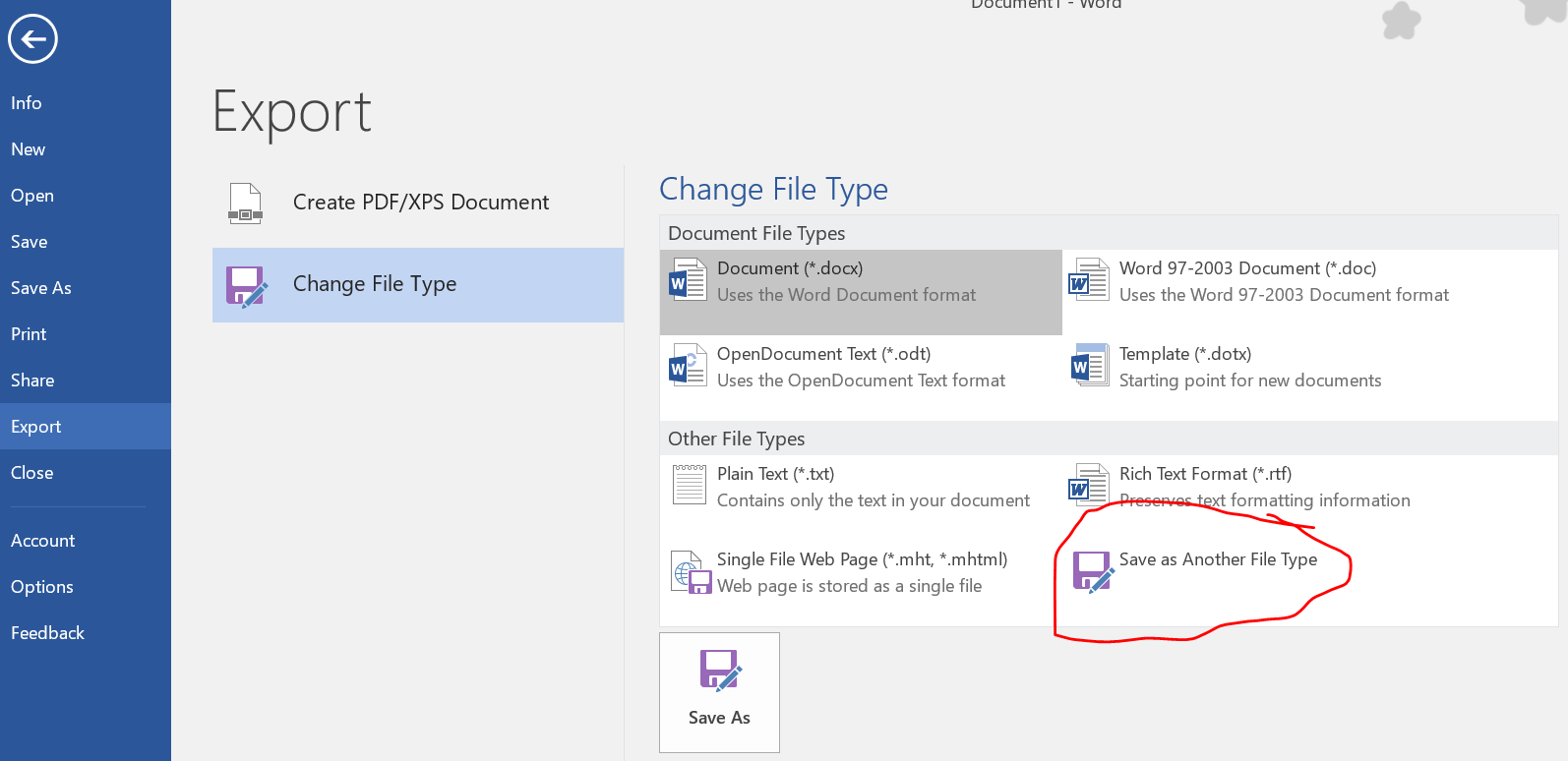
1. This is a great time to add other features that you think may help your model predict survivors better! Maybe the deck on which they stayed. Number of people they are responsible for. Etc… Read on the kaggle site for hints and to better understand the data!

A few tricks which might help:

=LEFT(<cell>,1) will return the first character in a cell

1 british pound in 1912 is now worth 87.66 US Dollars

## RESUME

When you feel you have added enough independent variables, save the excel spreadsheet as a CSV file. 

Machine generated alternative text:
Filename 
Save as type: CSV (Comma delimited) 
Authors: Andrew Moll 
le Folders 
Tags: 
Add a tag 
Tools 

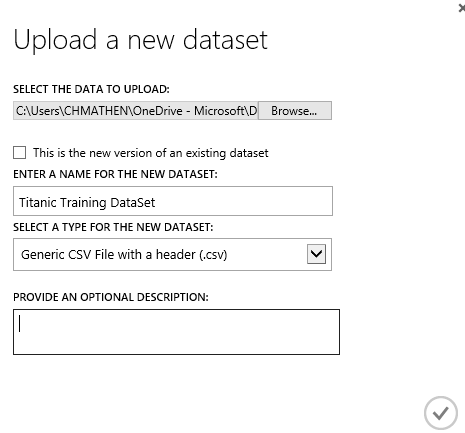
# Modeling

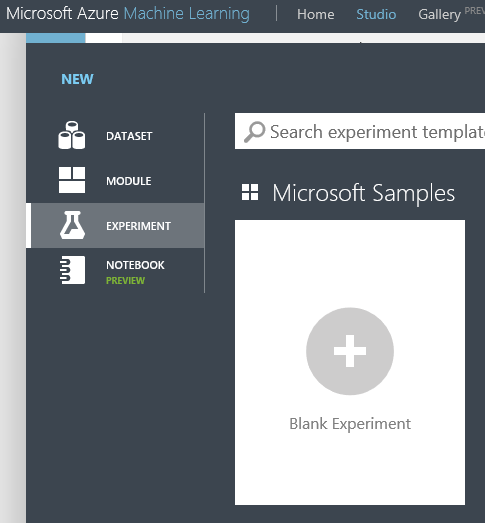
1. We're now going to upload this CSV as a dataset in Azure ML.

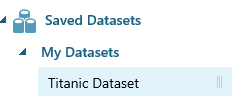
1. Navigate to <https://studio.azureml.net/> and login using your LiveID.
   1. If you run into issues try opening an inprivate browser session, Cntrl+Shift+N in Chrome.
2. On the bottom right side there will be a plus sign. Press this and choose datasets => upload new

Machine generated alternative text:
NEW 
DATASET 
MODULE 
EXPERIMENT 
NOTEBOOK 
PREVIEW 
FROM 
LOCAL FILE 
Upload a new dataset from a local file 

1. Upload your new dataset, naming it appropriately. EX:"TitanicTrainingDataset". Hit the check box when you’ve filled out the form, to finish the upload process.



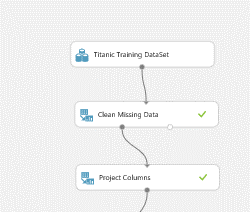
1. Next we'll create a new experiment from the same + symbol in the bottom left. Choose blank experiment and rename it appropriately.

1. Machine generated alternative text:
   I hope I survive! 
   TitanicTraindata 
   1 dataset (GenericCSV) Drag your newly added dataset to the Azure ML Canvas. This is on the right side, under Saved Datasets -> My Datasets

1. We need to clean missing data. Add a clean missing data module.

Machine generated alternative text:
TitanicTraindata 
Clean Missing Data 
2 

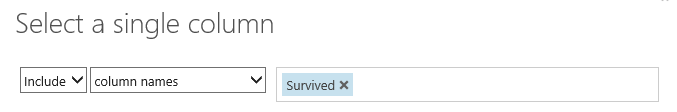
1. Next, we need to choose the right features. Add the "Project Columns" module. Here, you can choose your Independent Variables (or features)



* 1. Click "Launch column selector" to choose the columns that are the Independent variables that you think will be the best features to predict the Dependent Variable – also make sure you add your dependent variable as well.

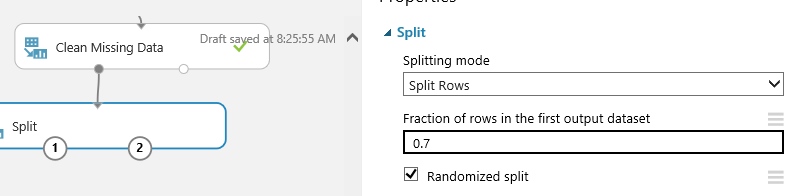
<Optional>

1. Machine generated alternative text:
   • 
   TitanicTraindata 
   Clean Missing Data 
   Filter Based Feature Selection There is another module type that we didn't cover in class, called a **Filter Based Feature Selection** which shows us which of our features have the most impact on our dependent variable.
   1. Launch the Column Selector to determine your Dependent Variable you want to use.



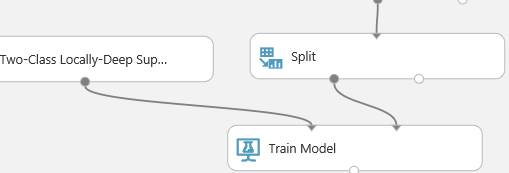
* 1. Change the number of desired features to more than 1, we chose 5.
  2. Press the play/run button.
  3. Machine generated alternative text:
     I hope I survive! ) Filter Based Feature Se ection ) Filtered dataset 
     rows 
     891 
     View as 
     columns 
     6 
     Survived 
     Pclass 
     3 
     Fare 
     7.25 
     71.2833 
     Parch 
     SibSp 
     Age 
     22 
     38 After running the experiment, you can click visualize and get the results below
  4. These results indicate that best independent variables to predict survived are Pclass, Fare, Parch, SibSp, and Age. What happens if you only use these as inputs to your model?

# </END OPTIONAL>

1. We'll also need to split the data set appropriately so that we have both training and testing data. Use a split module with 70% of the data being used to train the model and 30% to test

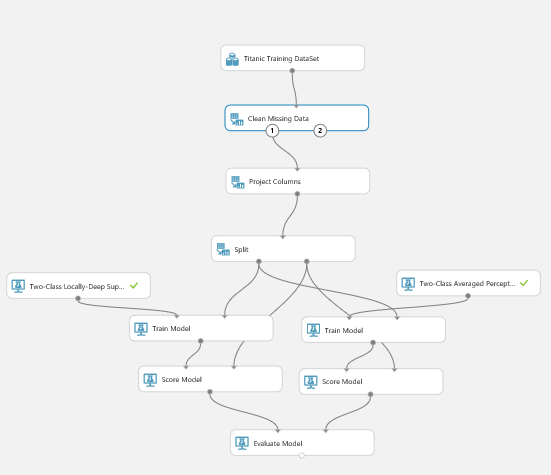
1. Next use the train model module, so drag that out as well, connecting the larger split data into train model. Don't forget to choose your dependent variable!

1. Now we know we're trying to solve a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ problem so we will add an algorithm from the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ drop down. Any will do to start! Add it into your experiment.
   1. See the pre-work page if you don’t remember



1. Add a score model module and connect the train model output to the top left node. Connect your holdback data, the 30% split, to the top right node of the score module.

1. Now recreate this structure (steps 11-13), reusing the data from the split module to train and score using another classification algorithm.

1. Now take the output of both score model and connect it to an evaluate model module.
2. At this point your canvas should look similar to the one below.

1. Press the play button on the bottom of the screen and wait for the model to finish running. Once it is finished running, click visualize on the evaluate model module and see how accurate your model is! Keep trying other algorithms until you are happy with your results.

1. Once you have decided which algorithm to save right click the bottom of the better train model module and select save as trained model.

Machine generated alternative text:
Two-Class Boosted Decision... 
Train Model 
split 
Download 
Save as Dataset 
Save as Trained Model 
Save as Transform 

AWESOME! You've just created your model

# Creating your Kaggle submission

1. Go back to Kaggle.com and download the test.csv file

Machine generated alternative text:
Competition Details » 
Data Files 
File Name 
train 
gendermodel 
genderclassmodel 
test 
Get the Data 
Make a submission 
Available Formats 
.csv (59.76 kb) 
.csv (3.18 kb) 
.csv (3.18 kb) 
.csv (27.96 kb) 

1. Open it in Excel and **add the same features that you had for your training set!**  This is super important. Your test data set must also have the same features as your training set.
   1. So, go back to Page 3 & 4 and repeat the process in excel

1. Upload your data set to AzureML using the same steps as before. This time, name is **Titanic Test Data**

1. Create a new experiment and name it Kaggle submission - <names>

1. Drag your newly saved model and uploaded dataset onto the canvas.

Machine generated alternative text:
Titanic Experiment 
Search experiment items 
Saved Datasets 
Titanic Dec 
Trained Models 
calhackathon 
Titanic Dec 
TitanicPre 
TitanicTest-pre.csv 

1. Use the score model module to score your test data set against your trained model. Your output will include your data as well as your scored labels. Kaggle expects your submission to be of the format

|  |  |
| --- | --- |
| PassengerID | Survived |
| 123 | 0 |
| 234 | 1 |

1. So we will need to change your output to match this format.

1. Let's use a project columns module to project only the columns needed for submission.

Machine generated alternative text:
Miment 
Titanic Dec 
TitanicTest-pre.csv 
Score Model 
Project Columns 
In draft 
Draft saved at PM 
Properties 
Project Columns 
Select columns 
Selected columns: 
Column names: 
Passengerld,Scored 
Labels 
Launch column selector 

1. Machine generated alternative text:
   Titanic Experiment 
   Titanic Dec 
   TitanicTest-pre.csv 
   Score Model 
   Project Columns 
   Metadata Editor 
   In draft 
   Draft saved at PM 
   Properties 
   Metadata Editor 
   Column 
   Selected columns: 
   Column names: Scored 
   Labels 
   Launch column selector 
   Data type 
   Unchanged 
   Categorical 
   Unchanged 
   Fields 
   Unchanged 
   New column names 
   Survived This gives us two columns PassengerID and ScoredLabels. Let's change the name of the second column to Survived using a metadata editor module.

Machine generated alternative text:
Titanic Dec 
TitanicTest-pre.csv 
Score Model 
Project Columns 
Metadata Editor 
Convert to CSV Lastly, convert your dataset to CSV and download it to your desktop!

# Submitting to Kaggle.com

1. Click "make a submission" on the left hand side of the page. Then upload your data set and see where you rank!
   1. Careful, you only get 10 submissions today and we want to see your best

Machine generated alternative text:
Dashboard 
Home 
Data 
Make a submission 
Information 
Description 
Evaluation 
Rules 
Prizes 
Frequently Asked Questions 
Further Reading / Watching 
Getting Started With Excel 
Getting Started With Python 
Getting Started With Pyth... 
Getting Started With Rand... 
New: Getting Started with R 
Submission Instructions 
Forum 
Scripts 
New Script 
Knowledge 3,373 teams 
Titanic: Machine Learning from Disaster 
Fri 28 sep 2012 
Competition Details » 
Make a submission 
Get the Data 
Make a submission 
Thu 31 Dec 2015 (2 months to go) 
File Format 
Your submission should be 
in CSV format. You can 
up oad this in a zip/gz/rar/7z 
archive if you prefer. 
# of Predictions 
We expect the so ution fi e to 
have 418 predictions. The 
fi e should have a header 
row. Please see the samp e 
submission file on the data 
page for an example of a 
valid submission. 
You have 10 entries today. This resets 23 hours from now (00:00 UTC). 
Click or drop your submission here 
Enter a brief description of this submission here. 

# Questions?

Andrew Moll [Andrew.moll@microsoft.com](mailto:Andrew.moll@microsoft.com)

Lauren Tran [Lauren.Tran@microsoft.com](mailto:Lauren.Tran@microsoft.com)

Christine Matheney [Christine.Matheney@Microsoft.com](mailto:Christine.Matheney@Microsoft.com)