Data Science Challenge
Conversion Model

Shirley Bao

June 8, 2017

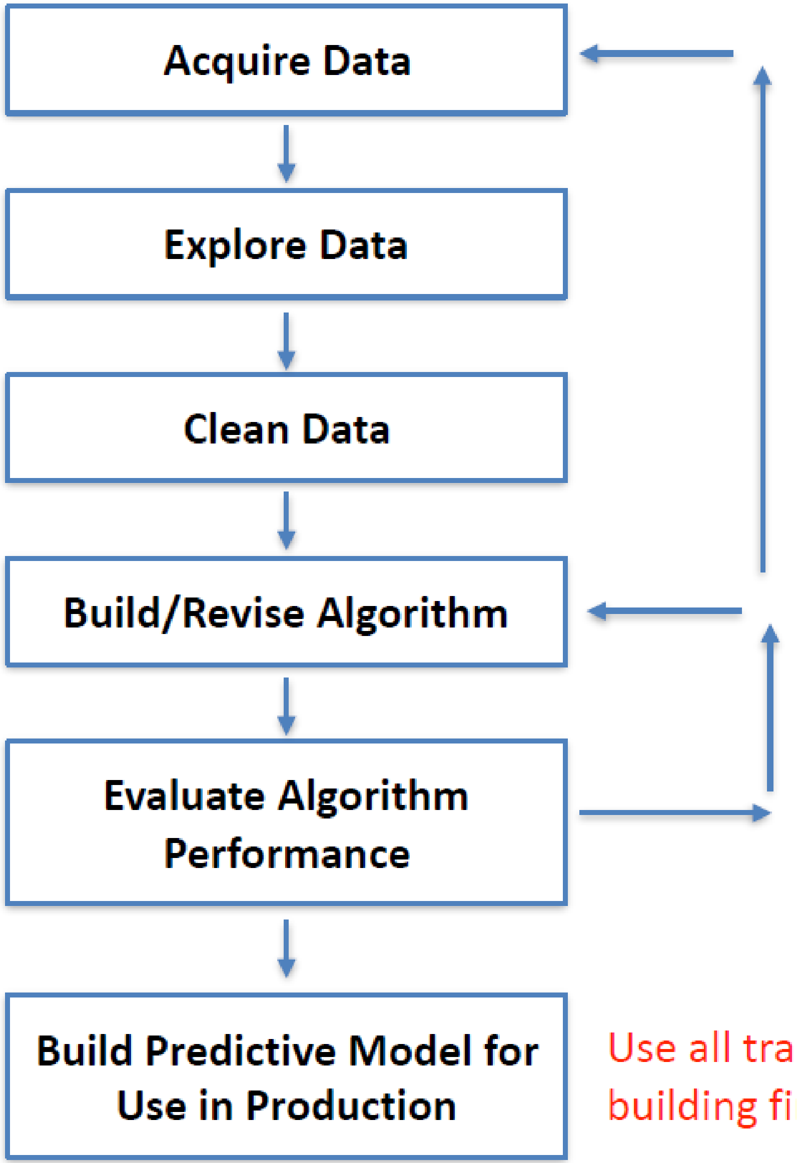
What you should expect during the next 2 hours

HANDS ON --- a case study approach to allow you to internalize the process of building useful data products and deriving actionable insights for business needs.

This session will be successful if you:

- Fully engage in the hands-on exercise and ask questions whenever you have any
- Contribute to our discussions and further apply the techniques in your day-to-day work

Data Science Process



Acquire Data

Explore Data

Clean Data

Build/Revise Algorithm

Evaluate Algorithm Performance

Build Predictive Model for Use in Production

Use all training data when building final model

Data Exploration

- Data types and missing values: Data.info()
- Basic Statistics: Data.describe().T
- Unique counts of values: pandas.Series**.value_counts**
- Distribution: Histograms
- Relationship with target: Scatter plots

Data Cleaning

- Erroneous Data
    - Small amount: remove
    - Treat as missing
    - Whether there is pattern in the error

- Missing Data
    - Imputation: mean, median, tree
    - Treat it as a separate category

- Non-numeric Categorical Data
    - Create dummy variables

Random Forest

- Ensemble algorithm (mix of many models)
- Collection of decision trees
- Evaluates predictions from many models and selects the most common classification
- Features are randomly selected for each decision tree
- Bagging (Bootstrap Aggregating) is also applied to each tree
    - Sample of the data set is used for training
    - Samples are drawn with replacement

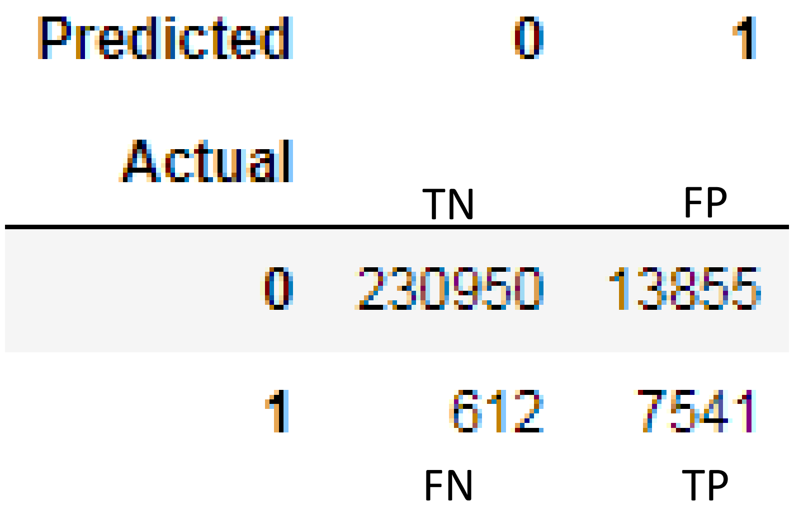
Why Random Forest Is Popular?

- Add all your data and algorithm will prioritize
- Not susceptible to overfitting
- Doesn’t require normalization
- Solid performance in wide range of applications

Classifiers Accuracy and Performance

- Confusion matrix: misclassification rate (MR), TP, TN, FP, FN, sensitivity and specificity
- ROC curves/AUC statistic
- Lift/Gains table and chart

Confusion Matrix



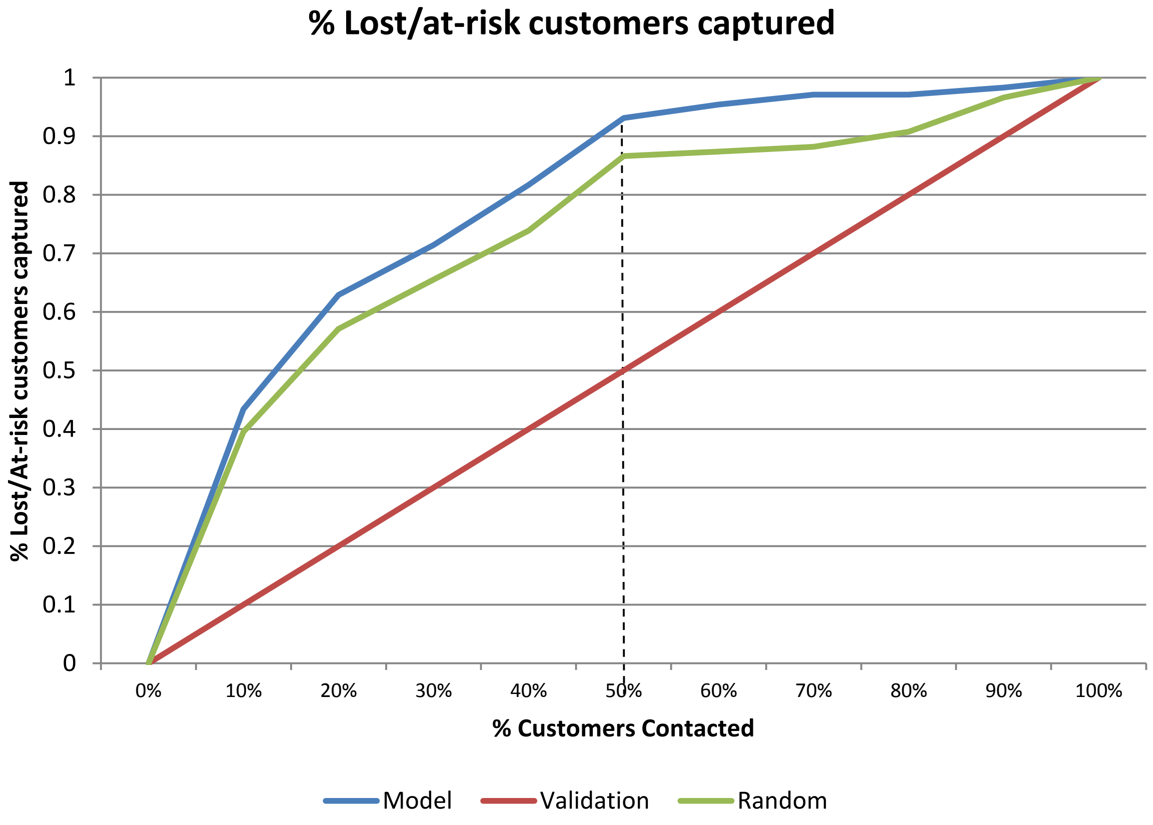
Predicted	0	1
Actual		
	TN	FP
0	230950	13855
1	612	7541
	FN	TP

N = 252,958

- TN = 230,950
- FP = 13,855
- FN = 612
- TP = 7541
- Misclassification Rate = (13855+612)/252958 = 5.7%
- TPR (sensitivity)= 7541/(7541+612) = 92%
- FPR = 13855/(230950+13855) = 5.7%
- Specificity = 1- FPR = 94%

Cumulative Lift/Gains Chart Example

New score formula allows us to target abandoners more effectively



Using the new scoring formula, we can reach out to 93% of all lost and at-risk customers by contacting the top 50% scored customers in the modeling sample and can reach 87% in the validation sample.