PYTHON PREDICTIONS
A TOBANIA COMPANY

# Q&A session & feedback Titanic

# Process for predictive (and descriptive) analytics

**Project definition**

**Data preparation**

**Model building**

**Model validation**

**Model usage**

# Process for predictive (and descriptive) analytics

**Project definition**

**Data preparation**

**Model building**

**Model validation**



**Model usage**



Make sure to also focus on these

# Model validation

**Metrics**

- **Accuracy** – dependent on probability threshold (scikitlearn's default threshold is 0.5)
  - Specificity, true positive rate (recall), false positive rate
- **AUC** – Area-Under-ROC-Curve, threshold-independent metric

**Graphical**

- **Lift** – how much better is model performing than a random guess?
- **Cumulative response** - how much better is model performing than a random guess?
- **Cumulative gains** – how many members of the positve group are we reaching with selection?

# Process for predictive (and descriptive) analytics

**Project definition**

**Data preparation**

**Model building**

**Model validation**

**Model usage**

**Iterative process** to improve models, with improved **feature engineering** and feature preprocessing

# Feature selection

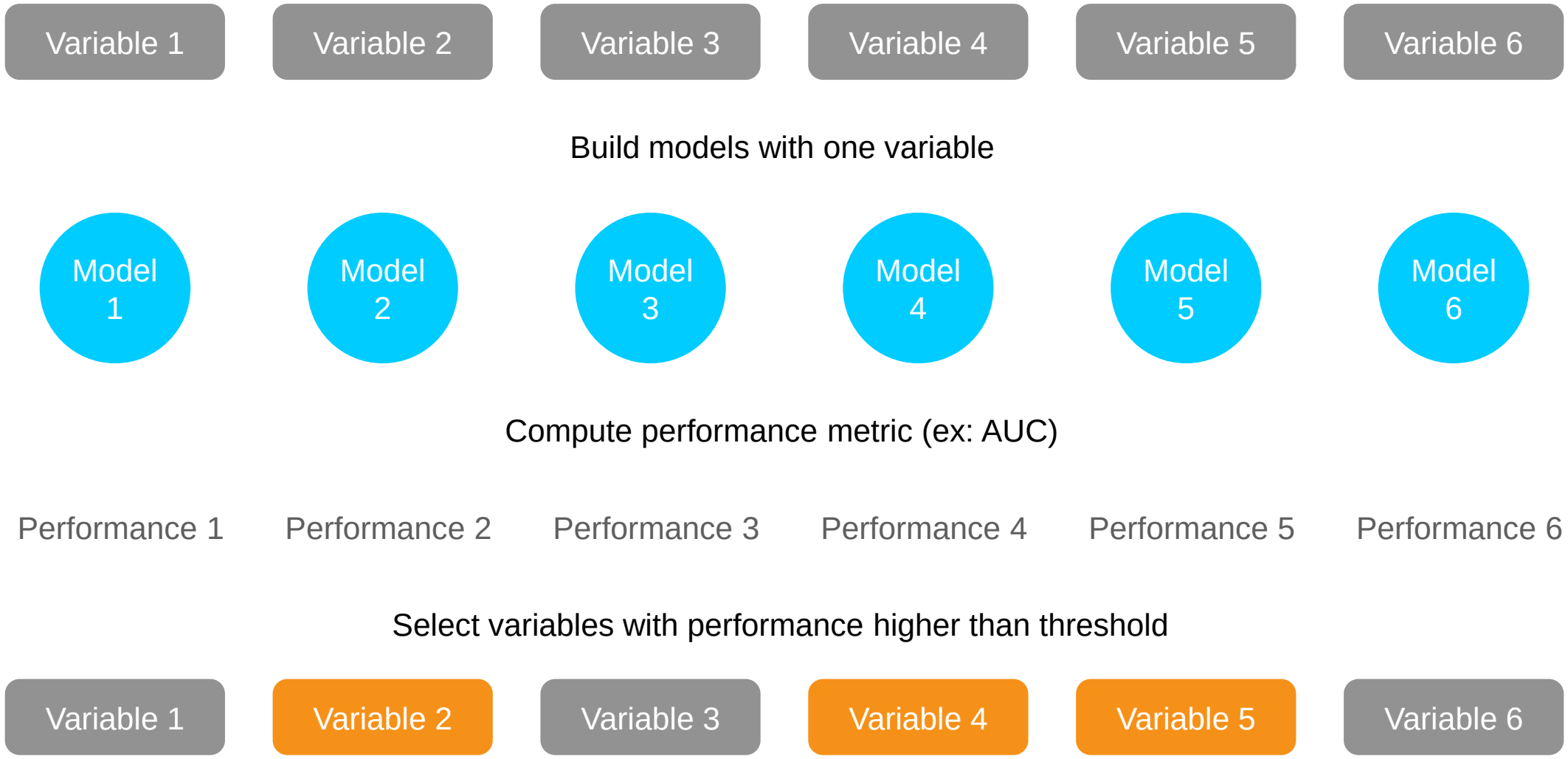
**Only select relevant features, because:**

- Overfitting
- Interpretability
- Efficiency (computational)
- Effort into industrializing more features

**Potential approaches:**

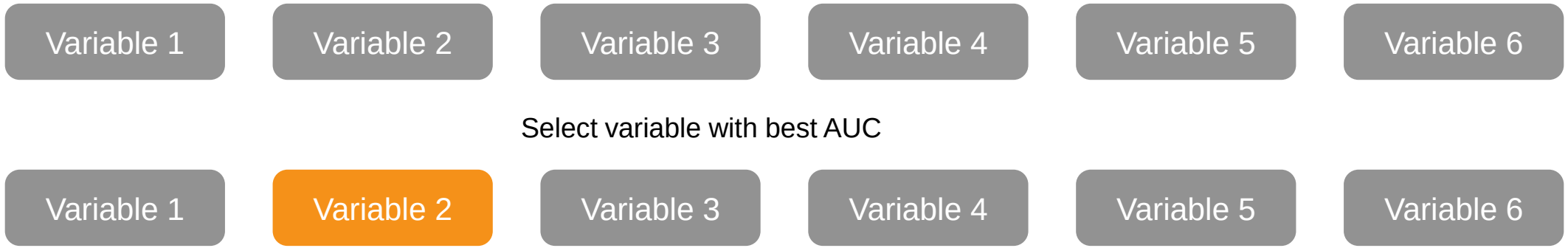
- Univariate feature selection
- Stepwise feature selection

# Univariate variable selection



# Stepwise forward feature selection

## Step 1 : Model with 1 variable



# Stepwise forward feature selection

## Step 2 : Model with 2 variables

# Stepwise forward feature selection

## Step 3 : Model with 3 variables

Variable 2
Variable 4
Variable 1

Variable 2
Variable 4
Variable 3

Variable 2
Variable 4
Variable 5

Variable 2
Variable 4
Variable 6

Select variables with best AUC

Variable 2
Variable 4
Variable 1

Variable 2
Variable 4
Variable 3

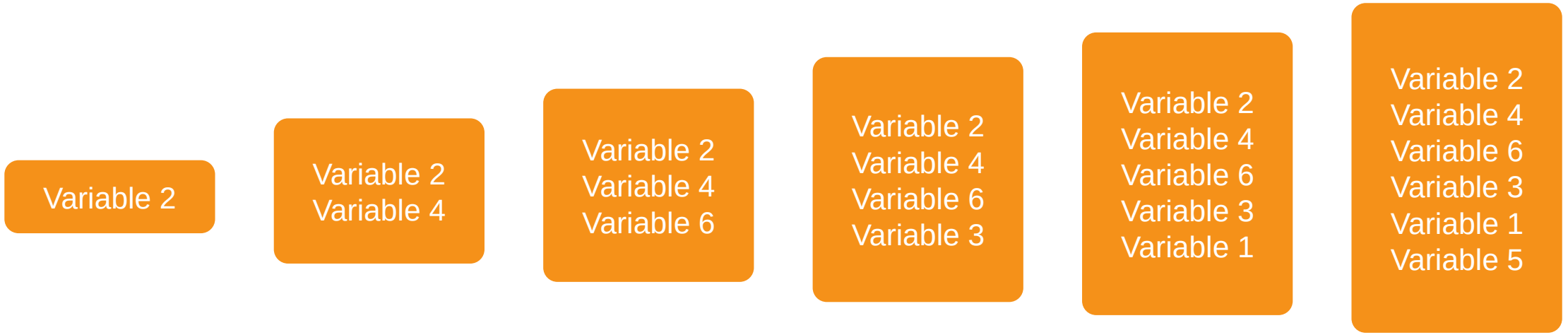
Variable 2
Variable 4
Variable 5

Variable 2
Variable 4
Variable 6

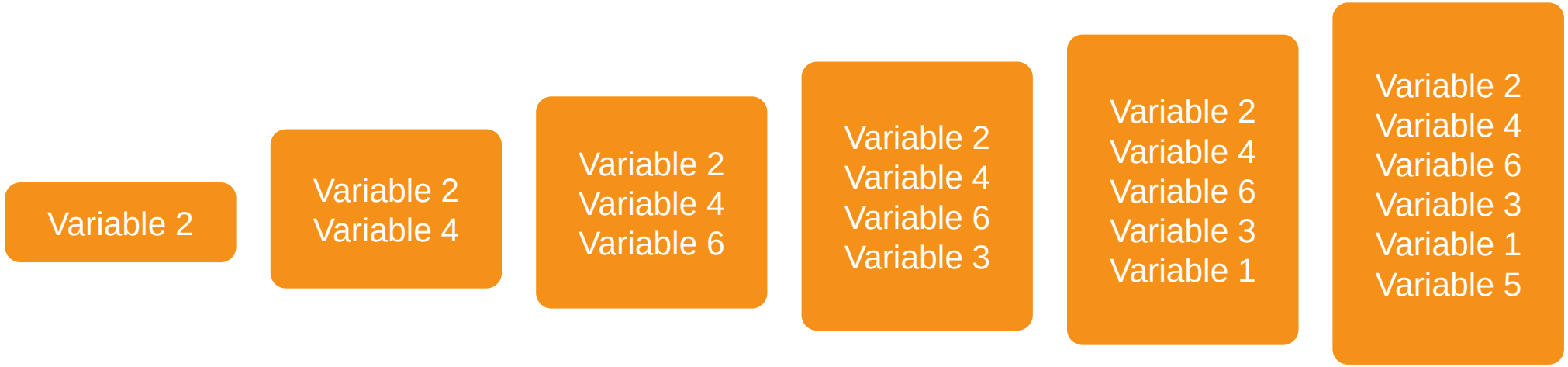
# Stepwise forward feature selection

Step N : Model with all variables

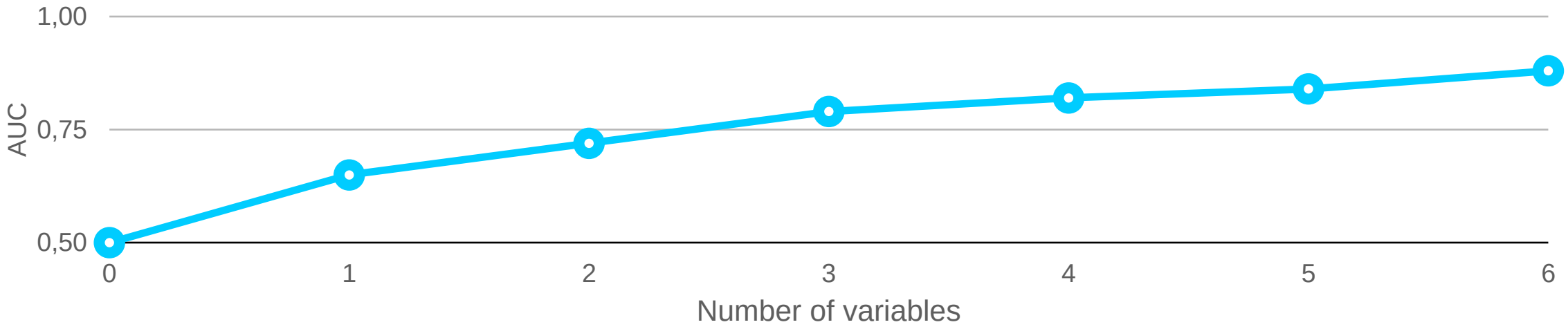
Result : N models, each with 1 additional variable



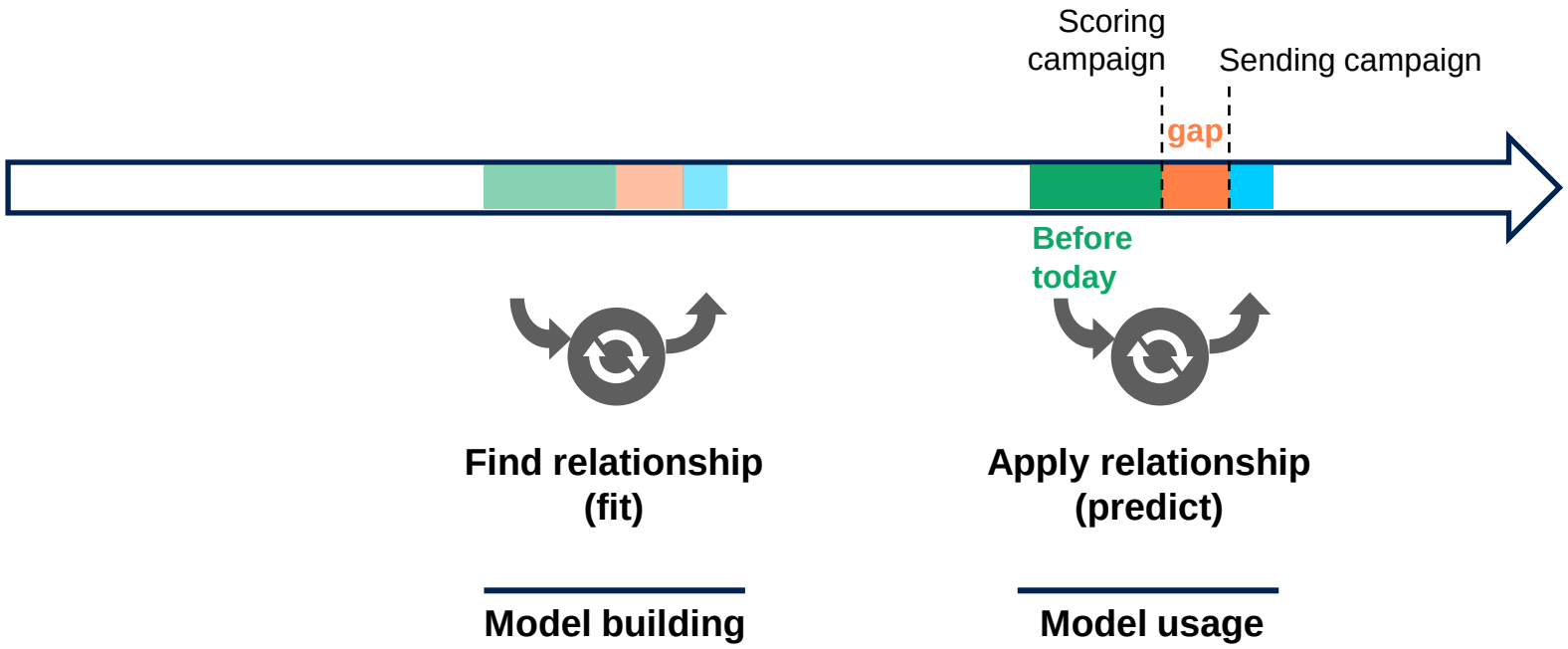
# Stepwise forward feature selection



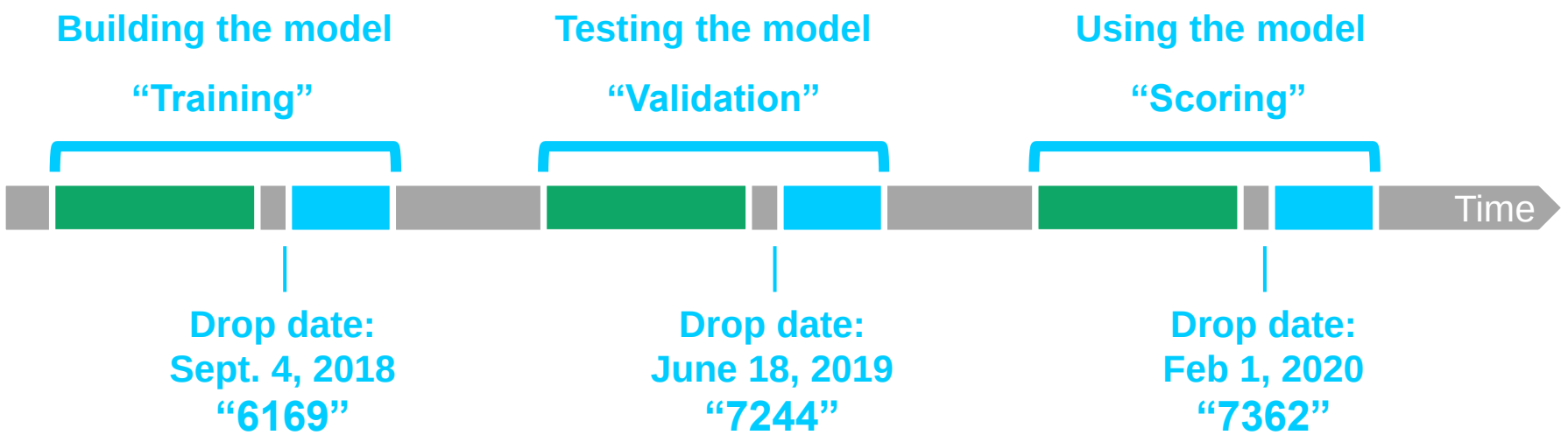

AUC = 0,65 AUC = 0,72 AUC = 0,79 AUC = 0,82 AUC = 0,84 AUC = 0,88



# Watch your timeline



# Watch your timeline




Business case assumes €0.80 per sent letter
Don’t consider the costs from the old campaigns.

# How can you push your model further?

1. Better features – do you capture all important things?
2. More training data
3. Double-check timeline
4. Different target (?)
5. Different model (linear vs. non-linear) & hyperparameter tuning

PYTHON PREDICTIONS
A TOBANIA COMPANY

PYTHON PREDICTIONS
A TOBANIA COMPANY

PYTHON PREDICTIONS
A TOBANIA COMPANY

PYTHON PREDICTIONS
A TOBANIA COMPANY

PYTHON PREDICTIONS
A TOBANIA COMPANY