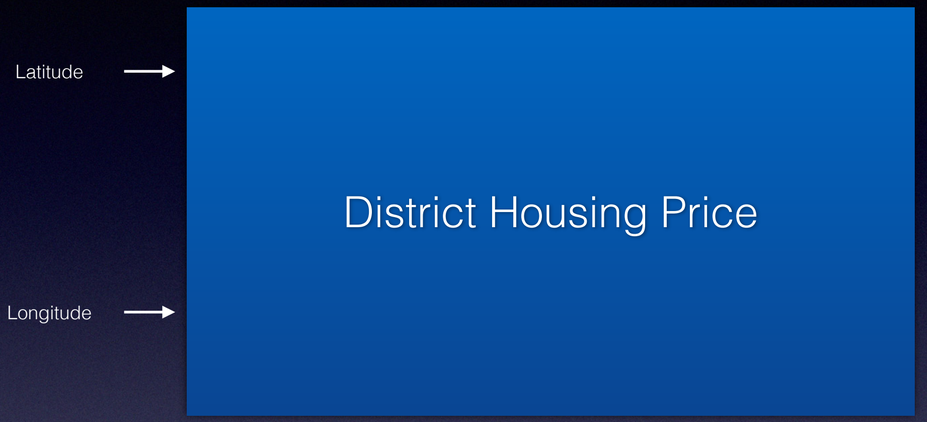
Chapter 2

End-to-End Machine Learning Project



Per Block Group(i.e. population of 600 to 3000 people)

District	Population	Median Income	Medium Price						

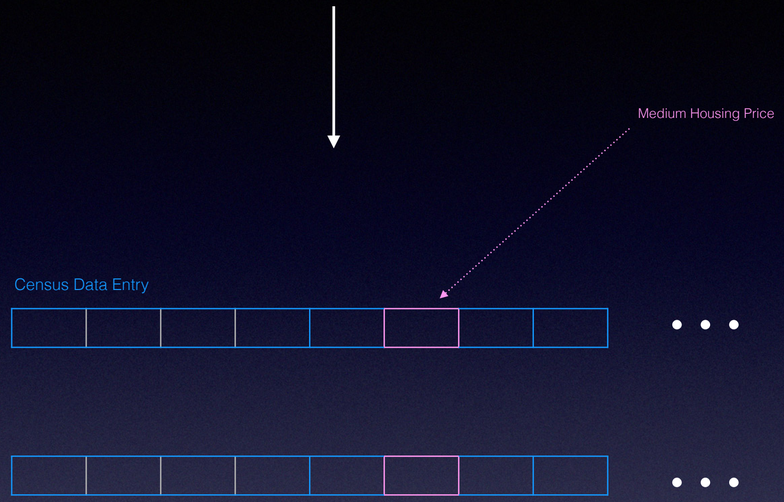
How will they use the model?

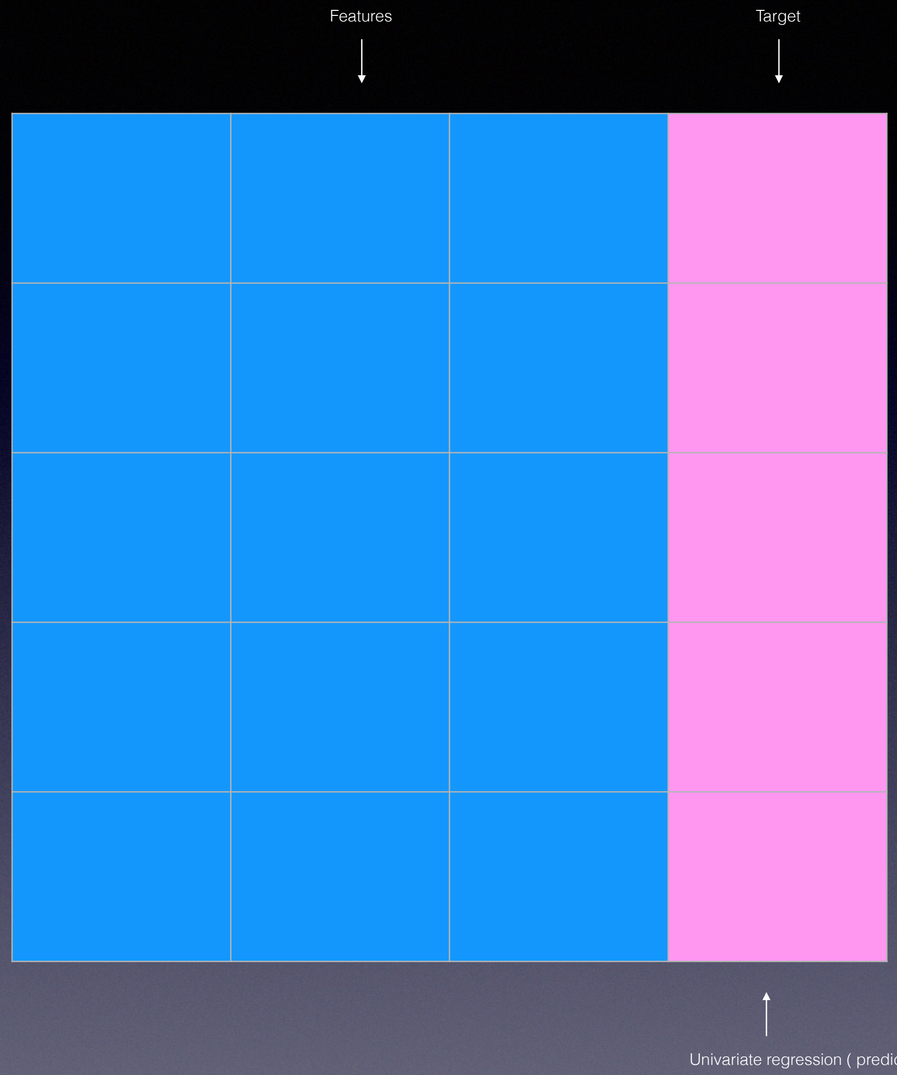
Model output will be fed into another system to determine whether it is worth investing in certain area.

How does company benefit from model?

It guides where we make investments, which directly impacts revenue at this company.

Use census data which provides medium housing price for districts in California




Features
Target
Univariate regression (predict single value)

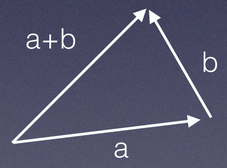
$$MAE(X, h) = \sqrt{\frac{1}{m}\Sigma_{i=1}^{m} | h(x^{(i)} - y^{(i)} |}$$

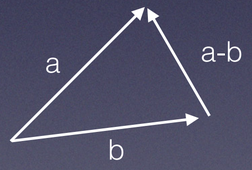
$$RSME(X, h) = \sqrt{\frac{1}{m}\Sigma_{i=1}^{m} (h(x^{(i)}) - y^{(i)})^2}$$

Cost functions measure distance of prediction vectors to target value vectors

RMSE - sensitive to outliers , use when outliers are negligible

MAE - not sensitive to outliers existing in dataset





Data split

crc32 redundacy per index

1696784233

2844319735

654825492

Index	…	…	…
0			
1			
2			
3			

⋮

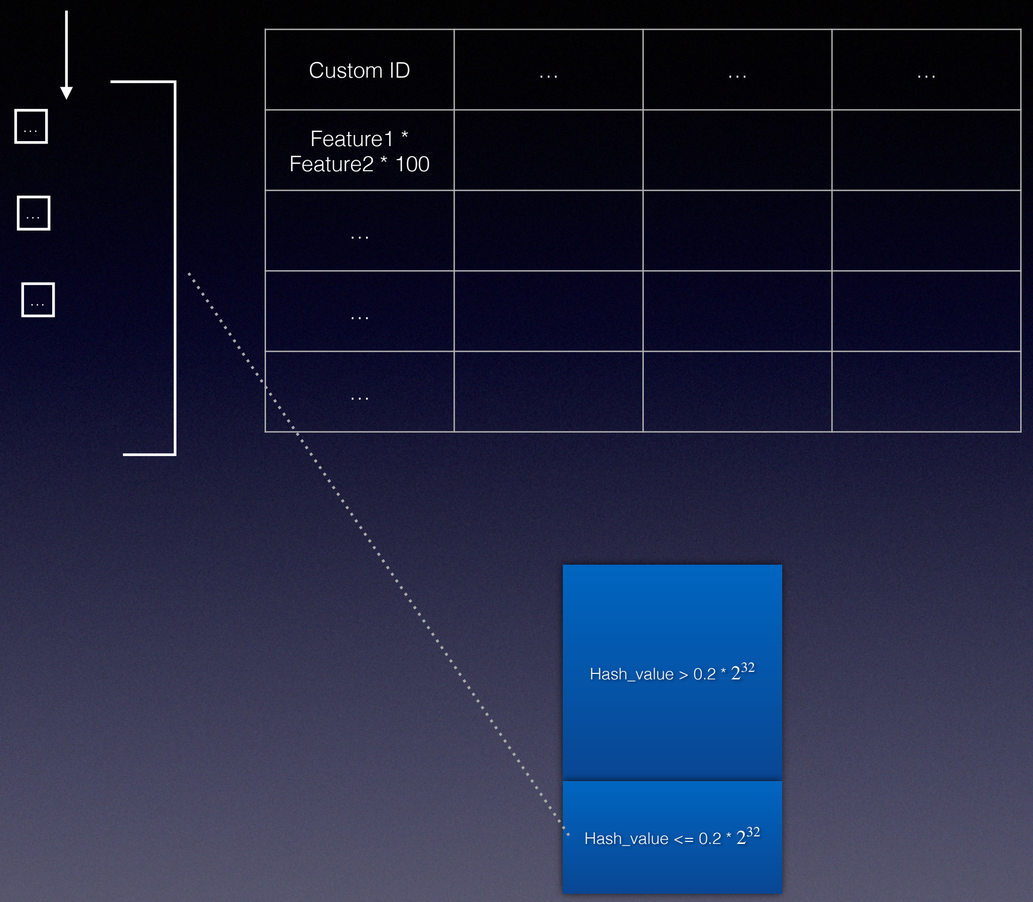
Hash_value > 0.2 * 2^{32}

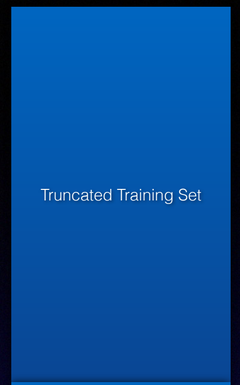
Hash_value <= 0.2 * 2^{32}

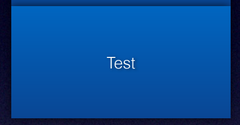
Data split

crc32 redundacy per index

Create custom ID using features if dataset gets deleted







Stratified sampling - sampling which guarantees the test is representative of overall population

Design Architect reveals the median income is very important or sensitive to predict median housing prices

Capture indices which represent the median income feature(i.e. attribute) when creating Test set.

Analyze Correlation to gain insights on data

Fill missing values in datasets using imputer

Encode Categorical Attributes

Ordinal Encoder

One Hot Encoder

Embedding

Low dimensional vector which is updated during training to support model performance

Custom Encoding

Replace country code with courtly population

Feature Scaling

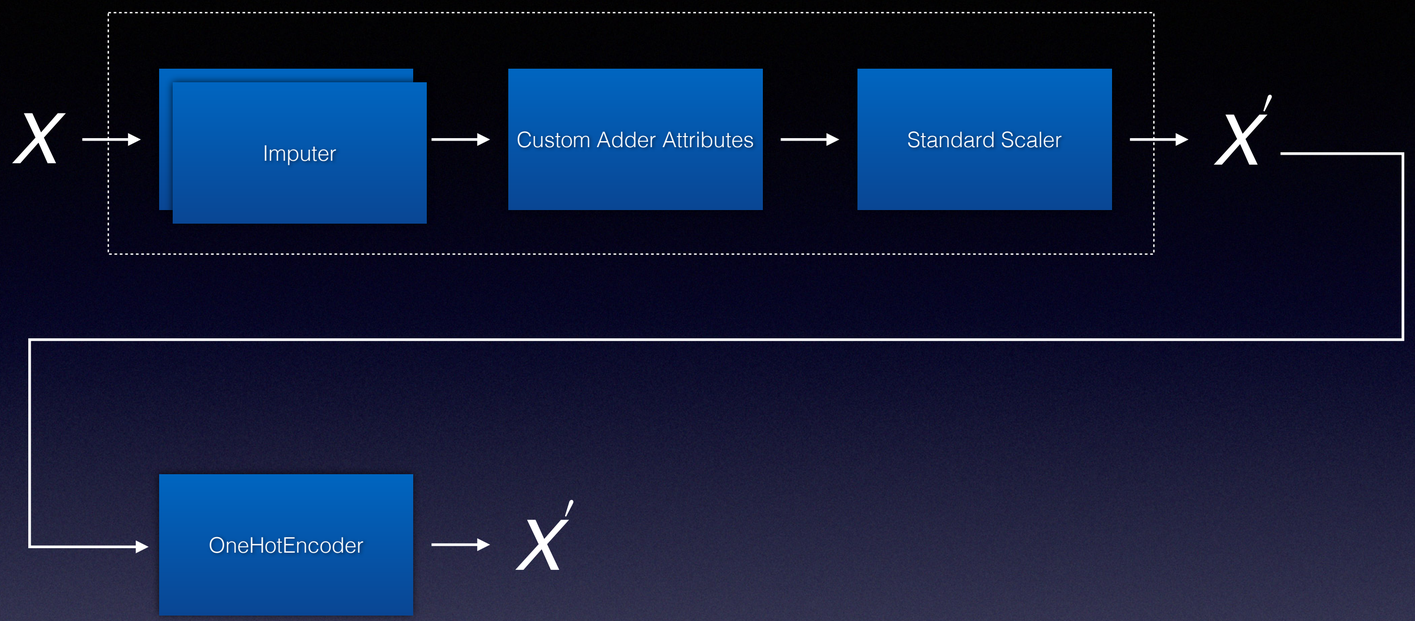
min-max

Values are rescaled to range from 0 to 1

standardization

Does not bound values to range like min-max.. This scaling option is not affected by outliers

Simple Transformation Pipelines

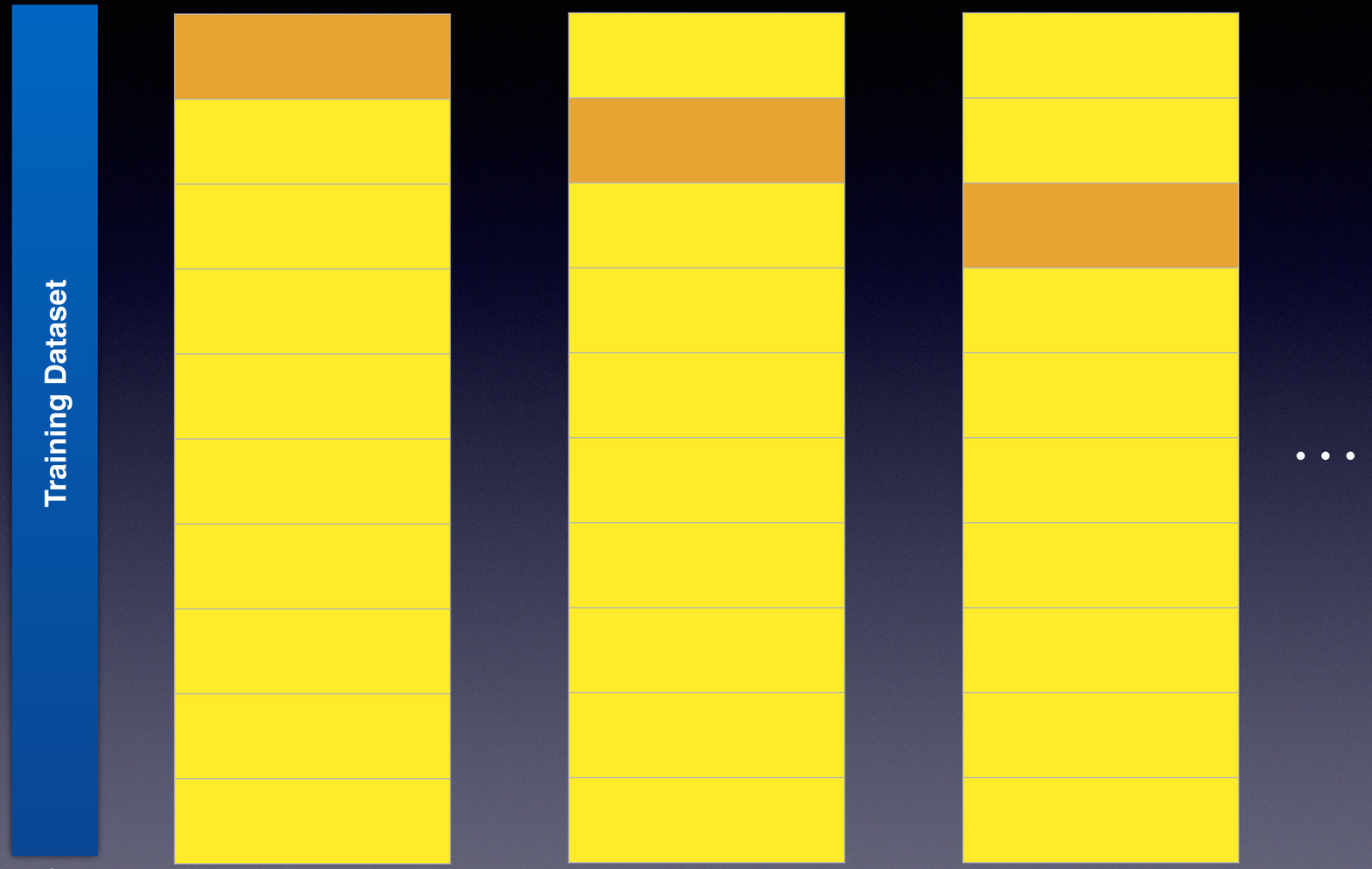


Cross Validation

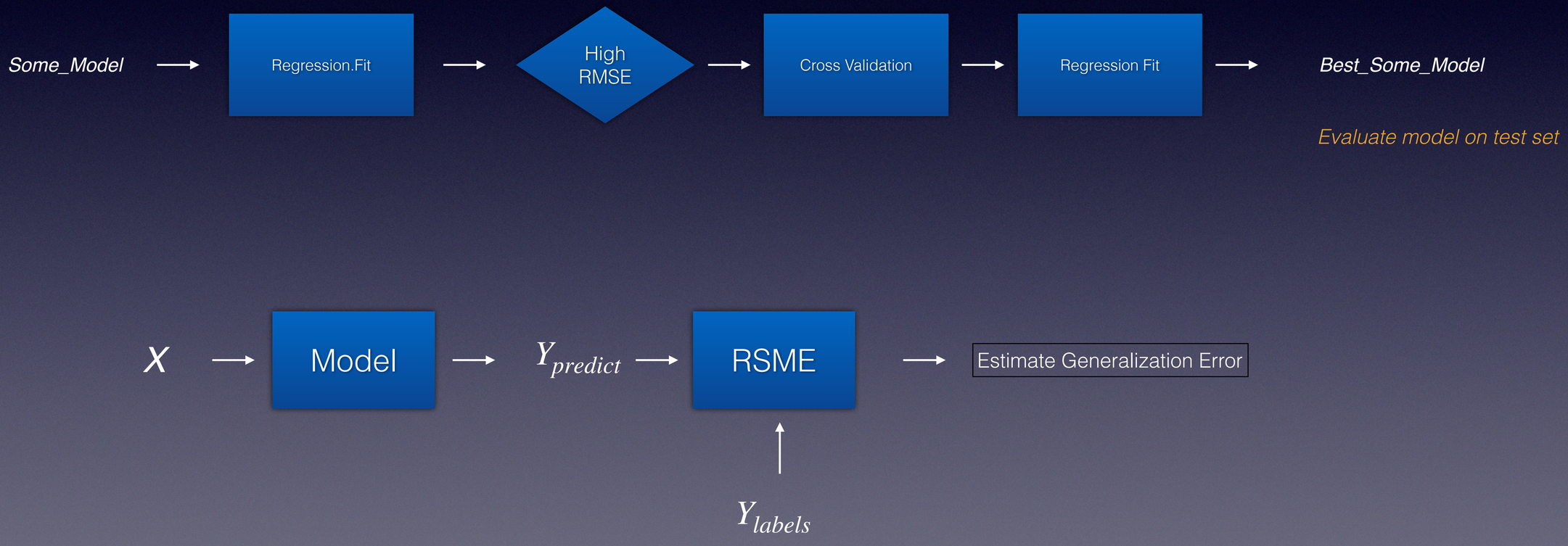
Train Set

Evaluation Set

CV = 10

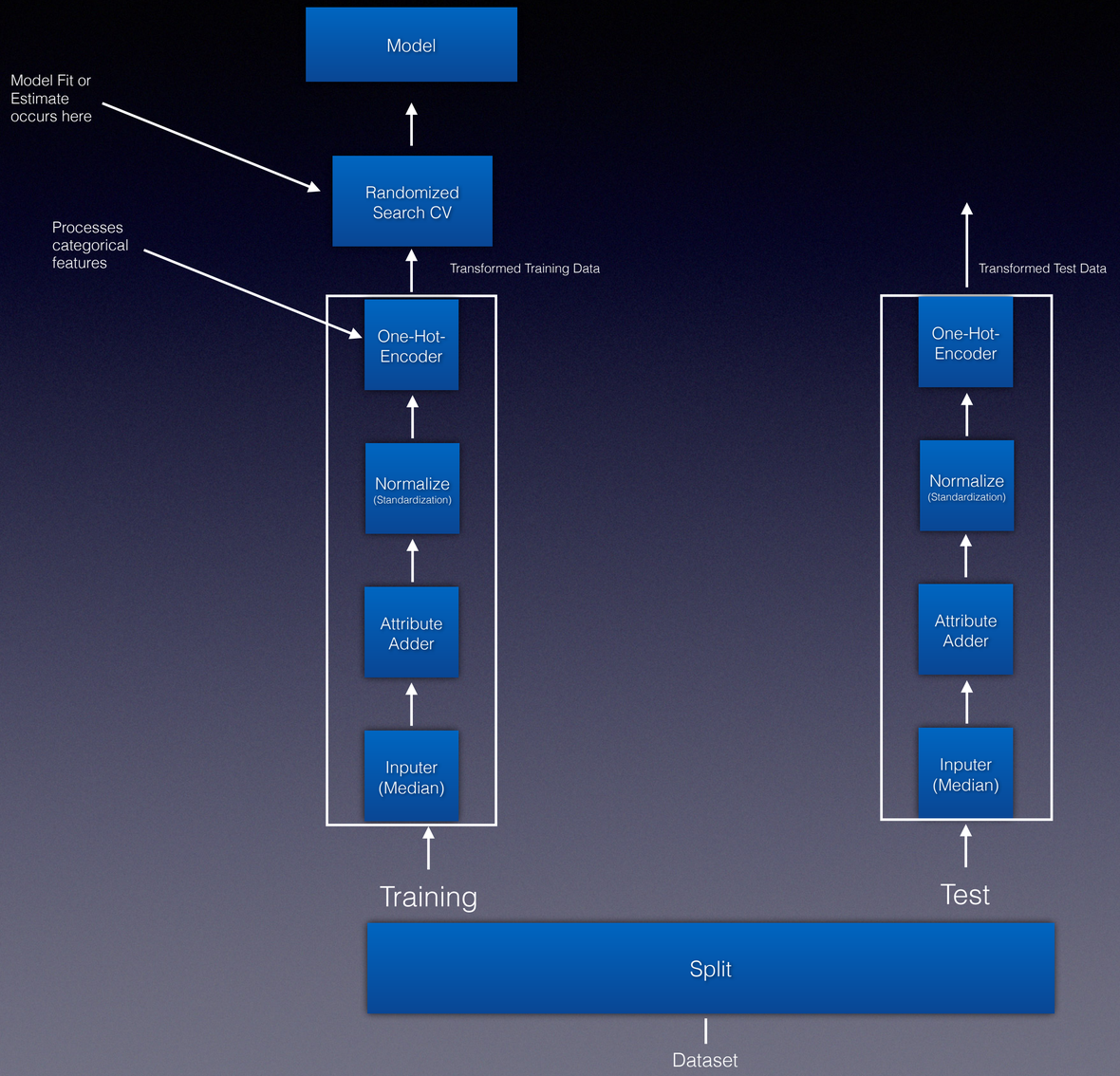


Training Model Score < Validation Score	Overfitting Model
Training Score = 0	Likely overfitting, run cross validation
Note	Cross validation estimates the performance of model on dataset ‘folds’, and estimates how precise the model is. Precision being its relation to other validation runs on the model


Some_Model
Regression.Fit
High
RMSE
Cross Validation
Regression Fit
Best_Some_Model
Evaluate model on test set
X
Model
$Y_{predict}$
RSME
Estimate Generalization Error
Y_{labels}

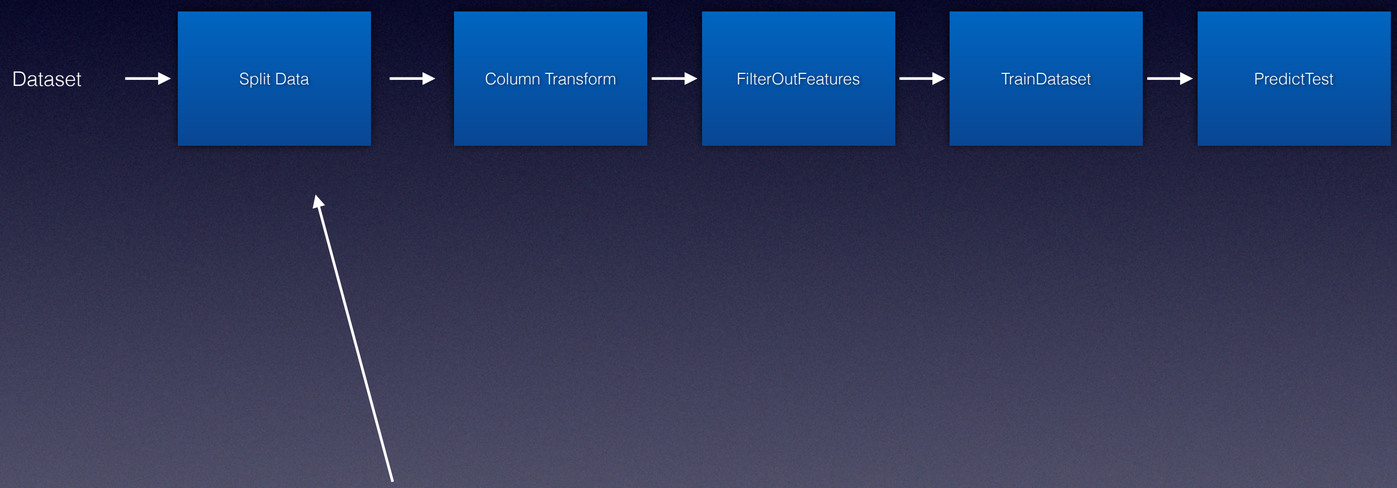
Estimate generalization error is estimated to lie between A(lower bound) and B(upper bound) with confidence level

Lets researcher know how precise the estimate is.


Model
Model Fit or Estimate occurs here
Randomized Search CV
Processes categorical features
Transformed Training Data
Transformed Test Data
One-Hot-Encoder
Normalize (Standardization)
Attribute Adder
Inputer (Median)
One-Hot-Encoder
Normalize (Standardization)
Attribute Adder
Inputer (Median)
Training
Test
Split
Dataset


Test Dataset
Model
Predictions
Statistics
Test Set Labels

Exercise Chapter 2



DerivedClass(BaseEstimator, TransformerMixin)