Hawaii Machine Learning Meetup

Introduction to Machine Learning in R

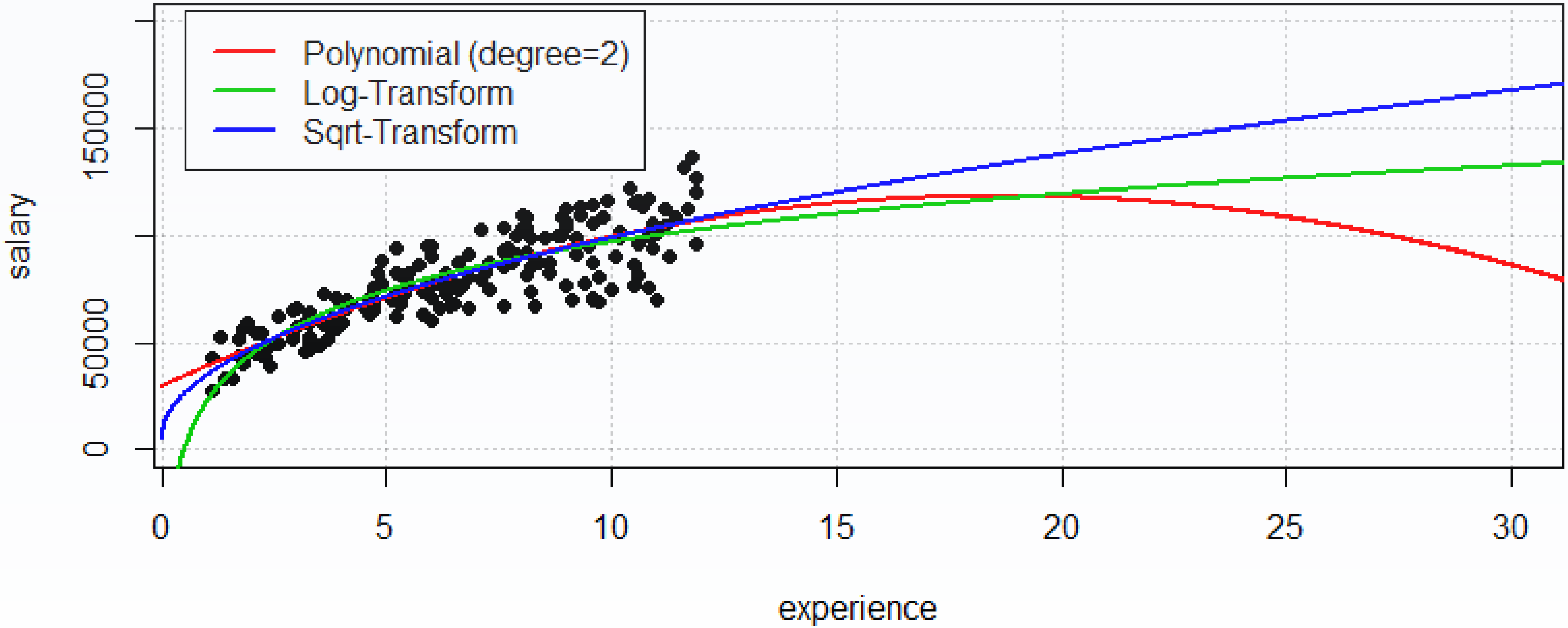


Table of Contents

- Recap
- Basics of R
- Exploratory Data Analysis
- Feature Engineering
- Model Selection
- Resources
- Hands-On Practice

Table of Contents

- Recap
- Basics of R
- Exploratory Data Analysis
- Feature Engineering
- Model Selection
- Resources
- Hands-On Practice

Recap

Introduction to Machine Learning in Python

- **Ingest**: import the data into a local data structure
- **Groom**: modify the data into some schema
- **Split**: break the data into a training set and a testing set
- **Select**: pick an algorithm appropriate for the data and the situation
- **Fit**: build a model of the data using the selected algorithm
- **Predict**: compute new results from the model
- **Display**: show a range of predictions from the model

Recap

Today’s Meetup

- **Exploratory Data Analysis**: gain insights
- **Feature Engineering**: incorporate insights and domain expertise
- **Model Selection and Overfitting**: determine which model is “best”

“Young man, in mathematics you don't understand things. You just get used to them.“

— John Von Neumann

Table of Contents

- Recap
- Basics of R
- Exploratory Data Analysis
- Feature Engineering
- Model Selection
- Resources
- Hands-On Practice

History

R is a language and environment for statistical computing and graphics

- Created by statisticians for statisticians

Ross Ihaka

Robert Gentleman

John Chambers

Basics of R

History

R is a language and environment for statistical computing and graphics

- Created by statisticians for statisticians
- Functional programming language

> *"To understand computations in R, two slogans are helpful:*
>
> - *Everything that exists is an object.*
> - *Everything that happens is a function call.“*
>
> *— John Chambers*

Table of Contents

- Recap
- Basics of R
- Exploratory Data Analysis
- Feature Engineering
- Model Selection
- Resources
- Hands-On Practice

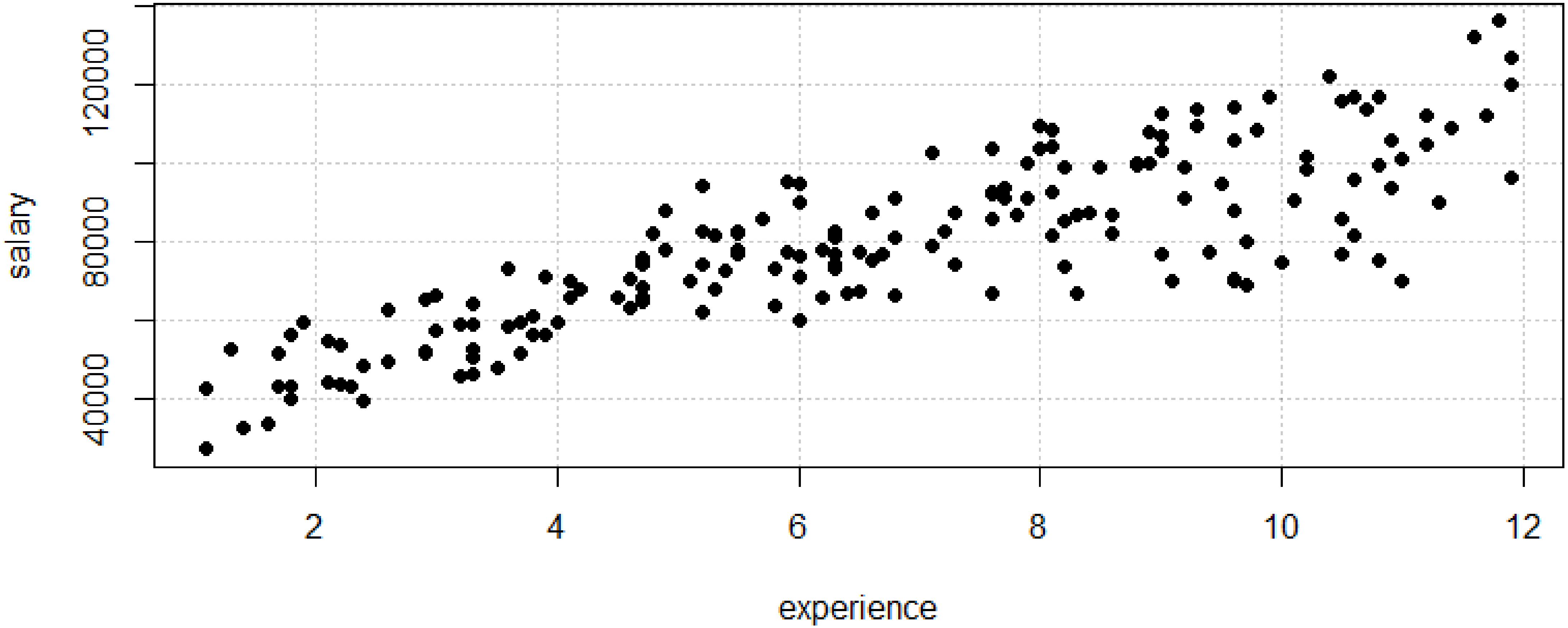
Exploratory Data Analysis

Exploratory data analysis is a process for understanding data

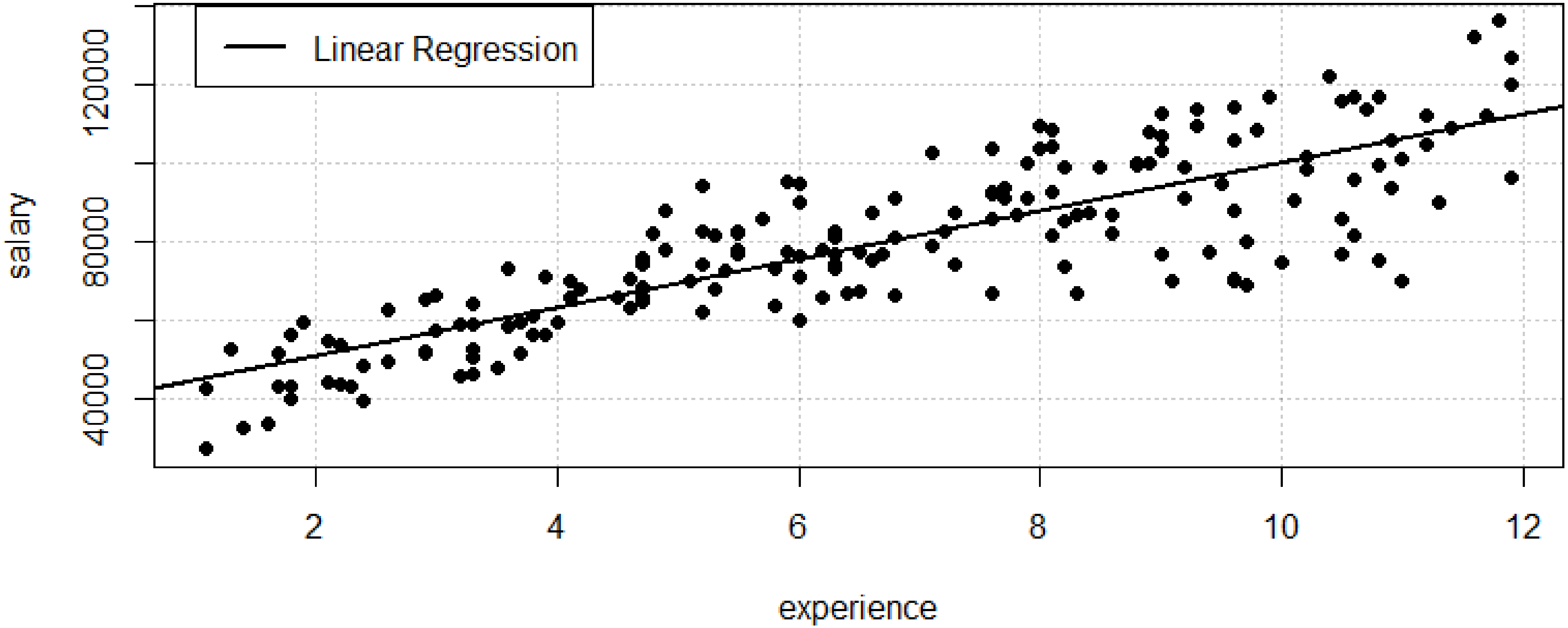
> *"Exploratory data analysis can never be the whole story,*
> *but nothing else can serve as the foundation stone."*
> *— John Tukey*

- Uncover underlying structure in a dataset
- Summarize characteristics of the dataset
- Maximize insight into a dataset

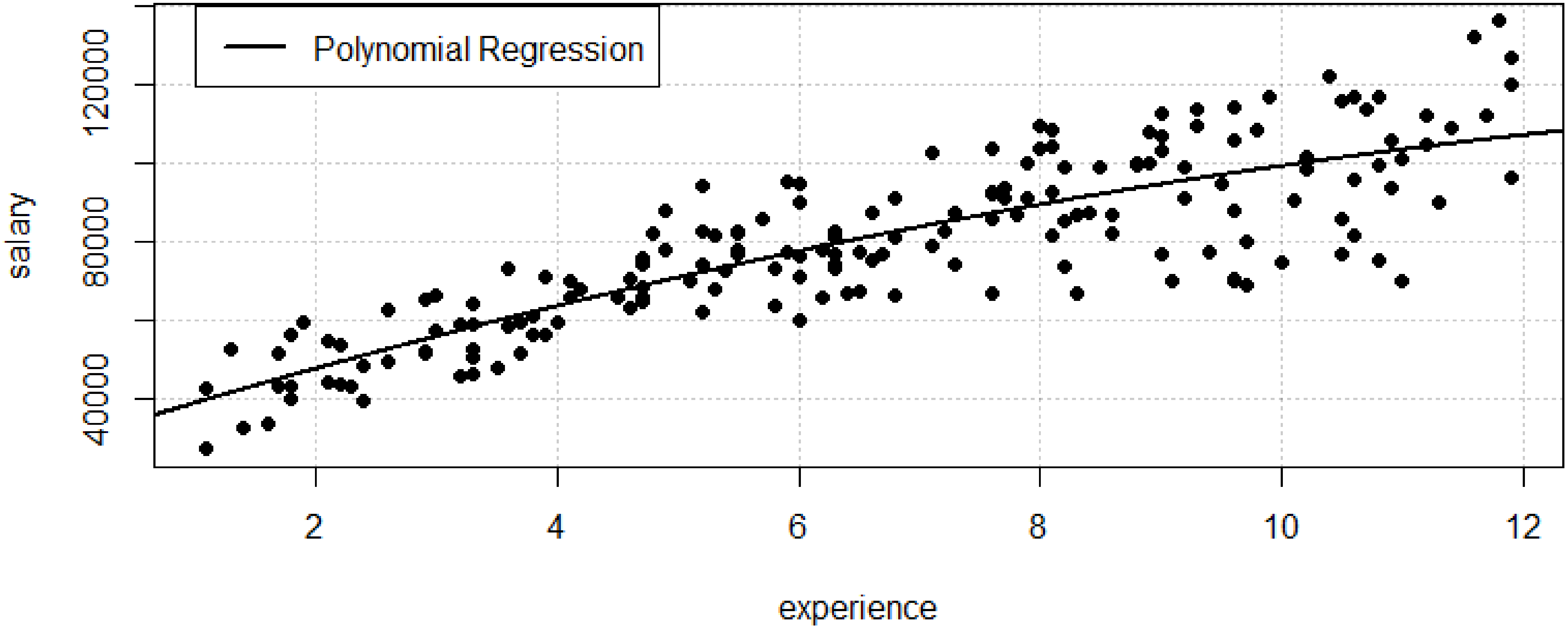
Exploratory Data Analysis



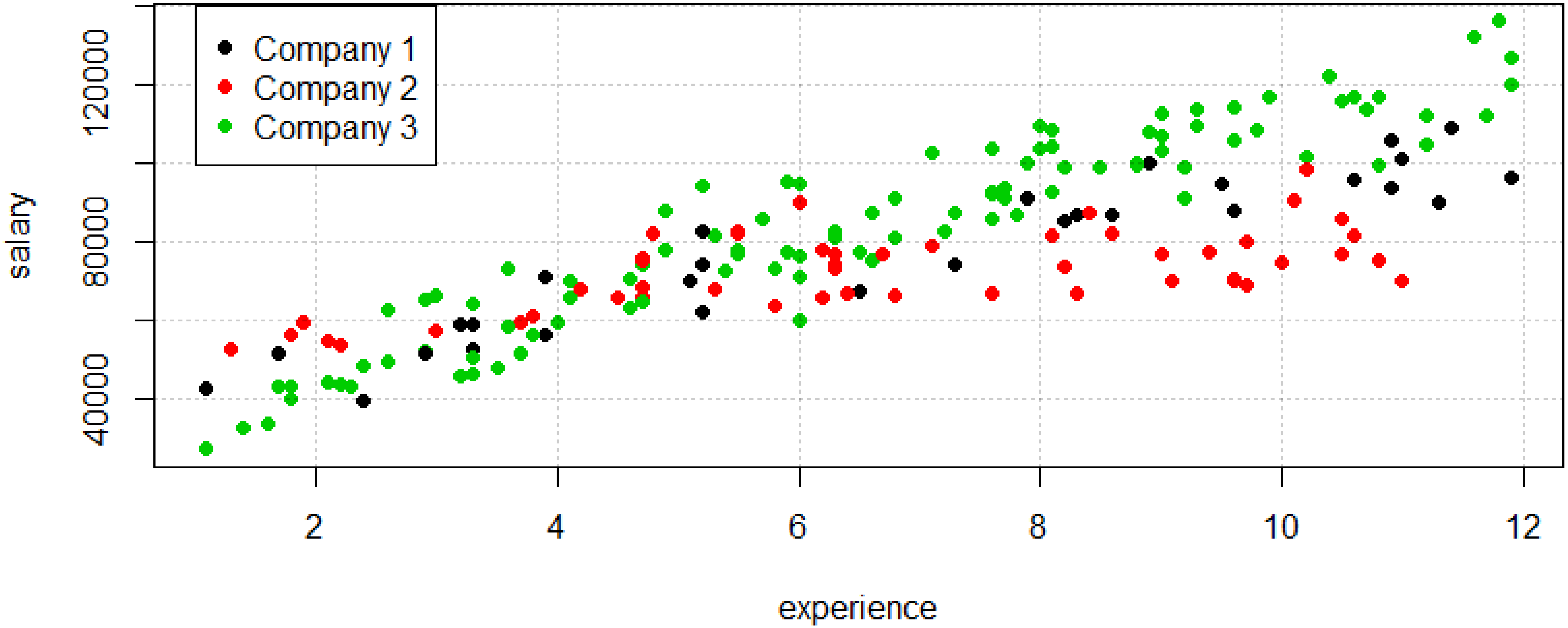
Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis

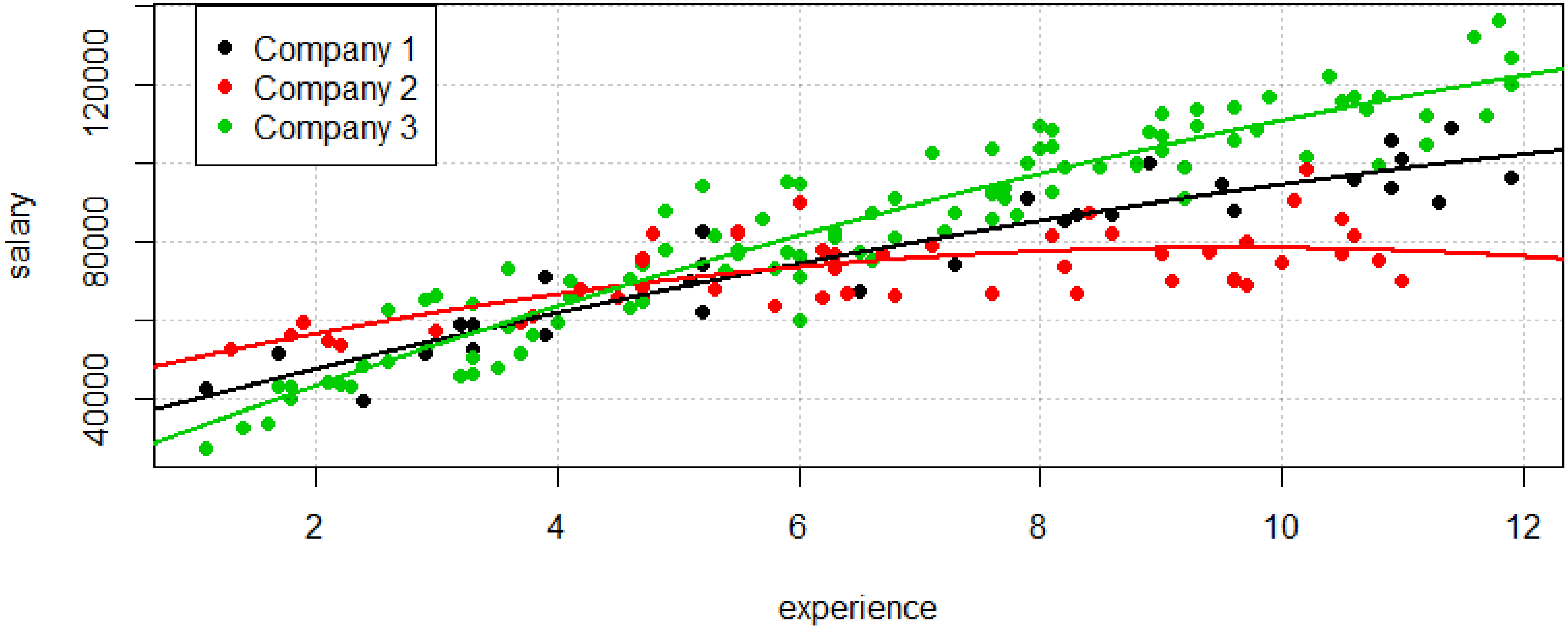


Table of Contents

- Recap
- Basics of R
- Exploratory Data Analysis
- Feature Engineering
- Model Selection
- Resources
- Hands-On Practice

Feature Engineering

Feature engineering is the process of creating new features

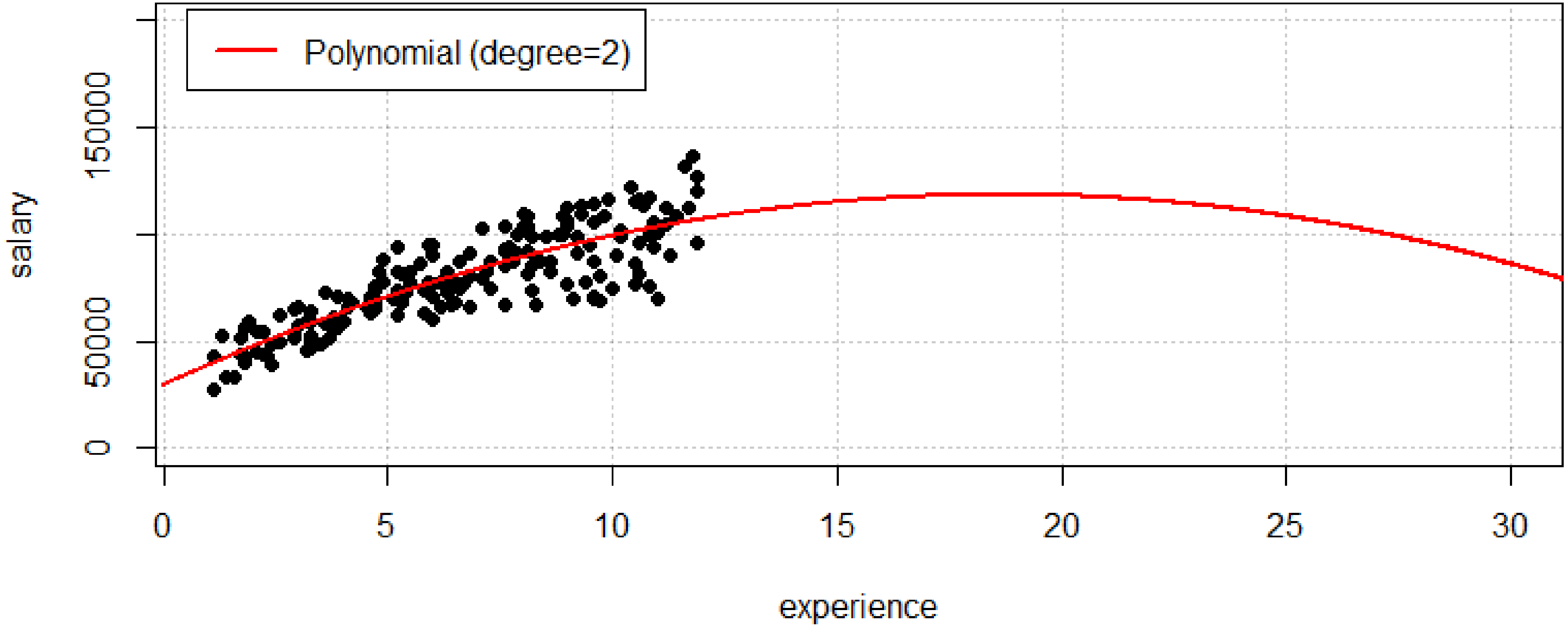
> *“Coming up with features is difficult, time-consuming, requires expert knowledge. "Applied machine learning" is basically feature engineering.”*
>
> *— Andrew Ng*

- Incorporates domain knowledge and intuition.
- Makes learning easier for the machine learning algorithm.

Feature Engineering

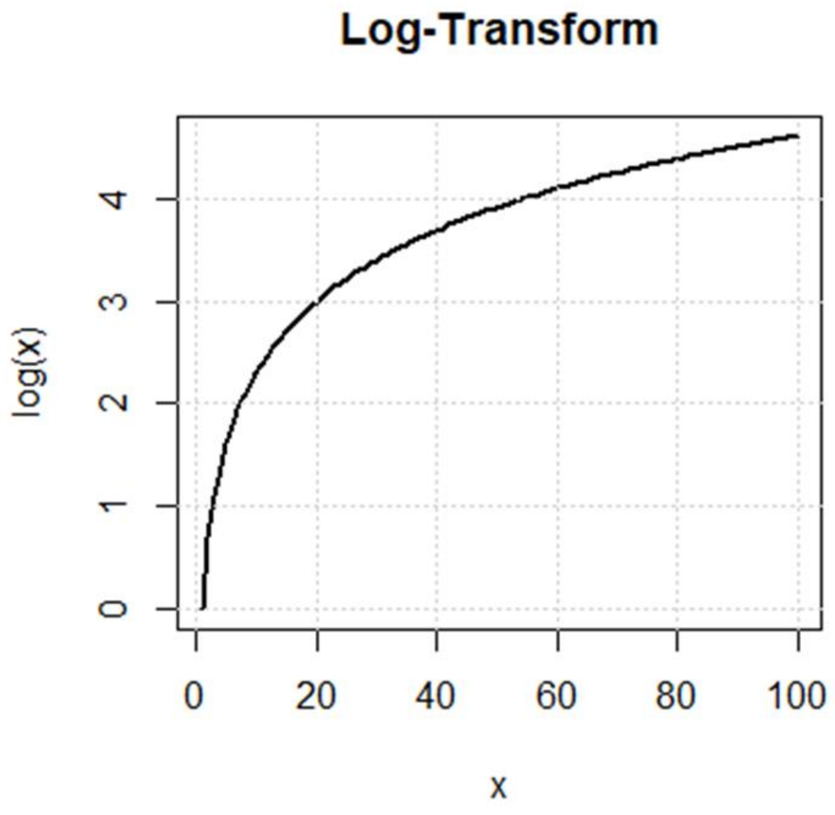
Monotonic Transformations

Salary should continuously increase with increasing experience.

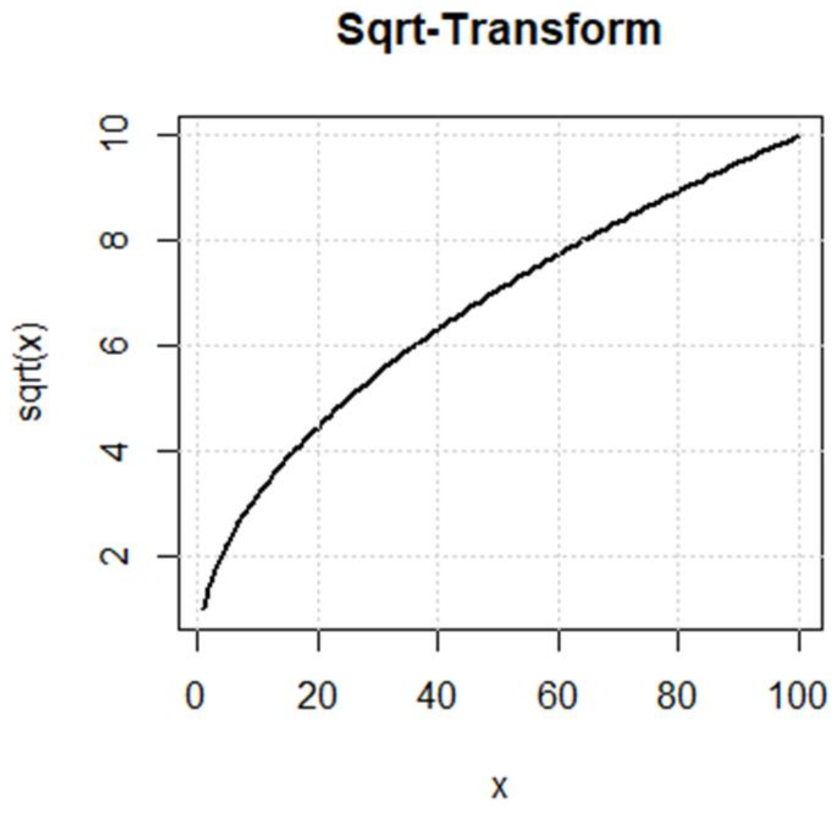


Monotonic Transformations

Salary should continuously increase with increasing experience.



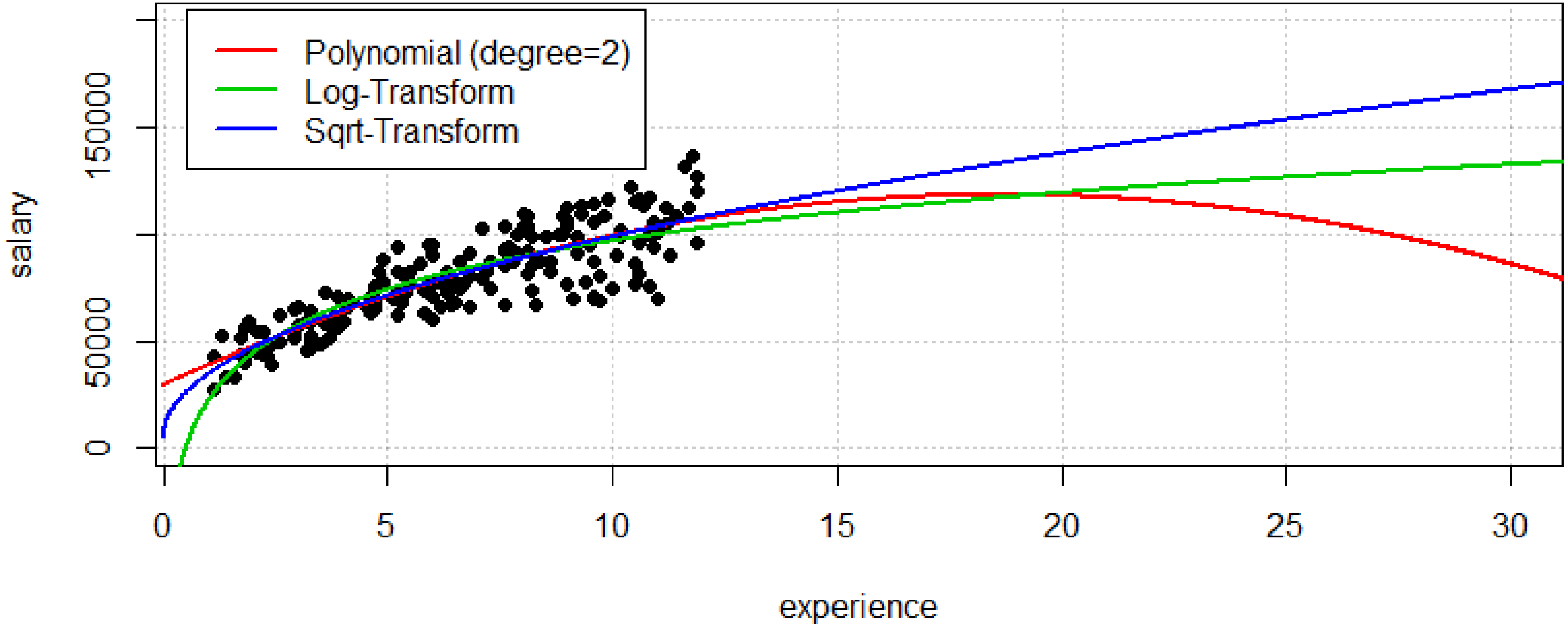
Log-Transform



Sqrt-Transform

Monotonic Transformations

Salary should continuously increase with increasing experience.



Feature Engineering

One-Hot-Encoding and Feature Interactions

$$\hat{y} = \sum_{i=1}^{3} I\{\text{company} = i\} \cdot \left(a_i \sqrt{\text{experience}} + b_i\right)$$

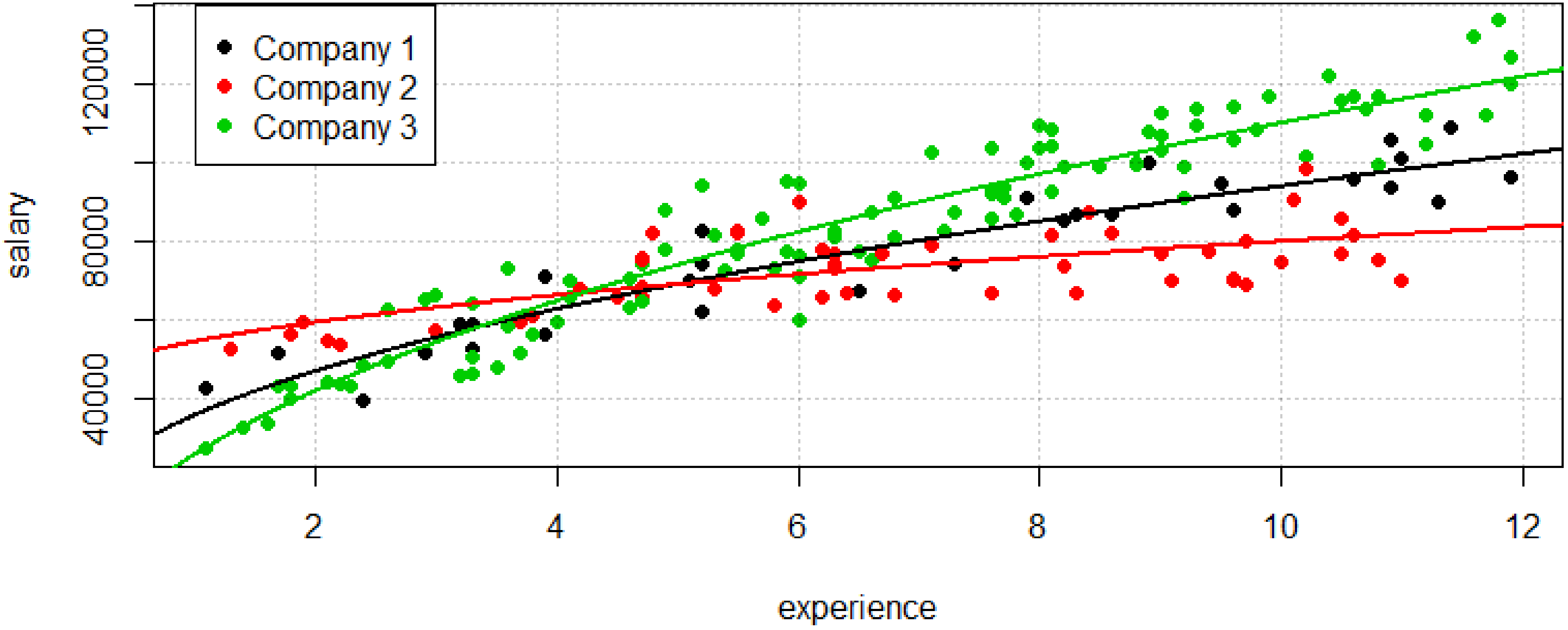


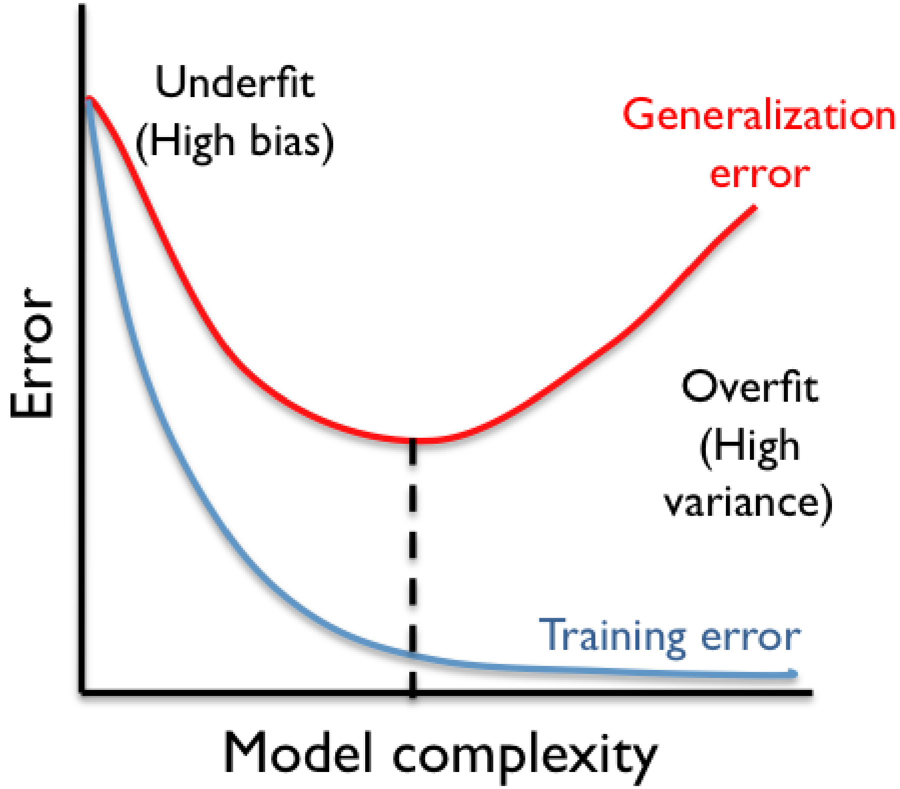
Table of Contents

- Recap
- Basics of R
- Exploratory Data Analysis
- Feature Engineering
- Model Selection
- Resources
- Hands-On Practice

Model Selection

Model selection addresses the following questions:

- How do we know which features to use?
- How do we know which model is “best”?
- What do we mean by “best”?



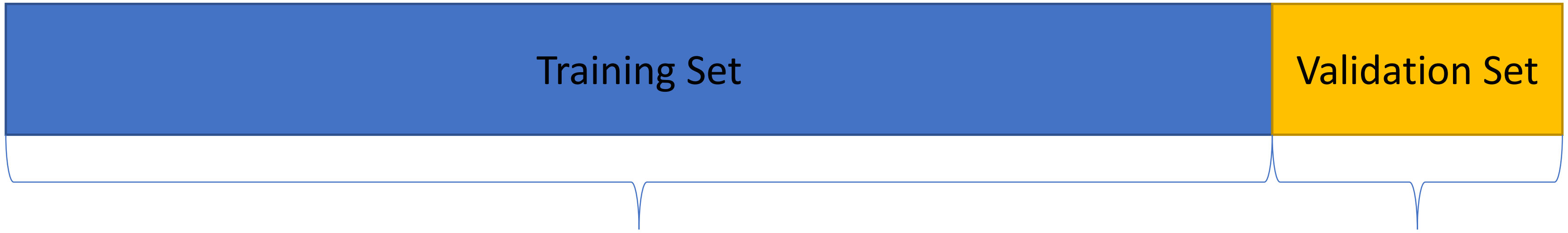
Training, Validation, and Generalization Error

- We fit a model to minimize training error.
- We evaluate a model using validation error.
- Our theoretical performance of a model is given by it’s generalization error.

Model Selection

Single Validation Set

- Partition the data into a training set and a validation set.
- Fit a model by minimizing training set error.
- Make predictions on the validation set.
- The validation error is an estimate of the generalization error.



Fit a model by minimizing error on the training set

Validation Error
≈ Generalization Error

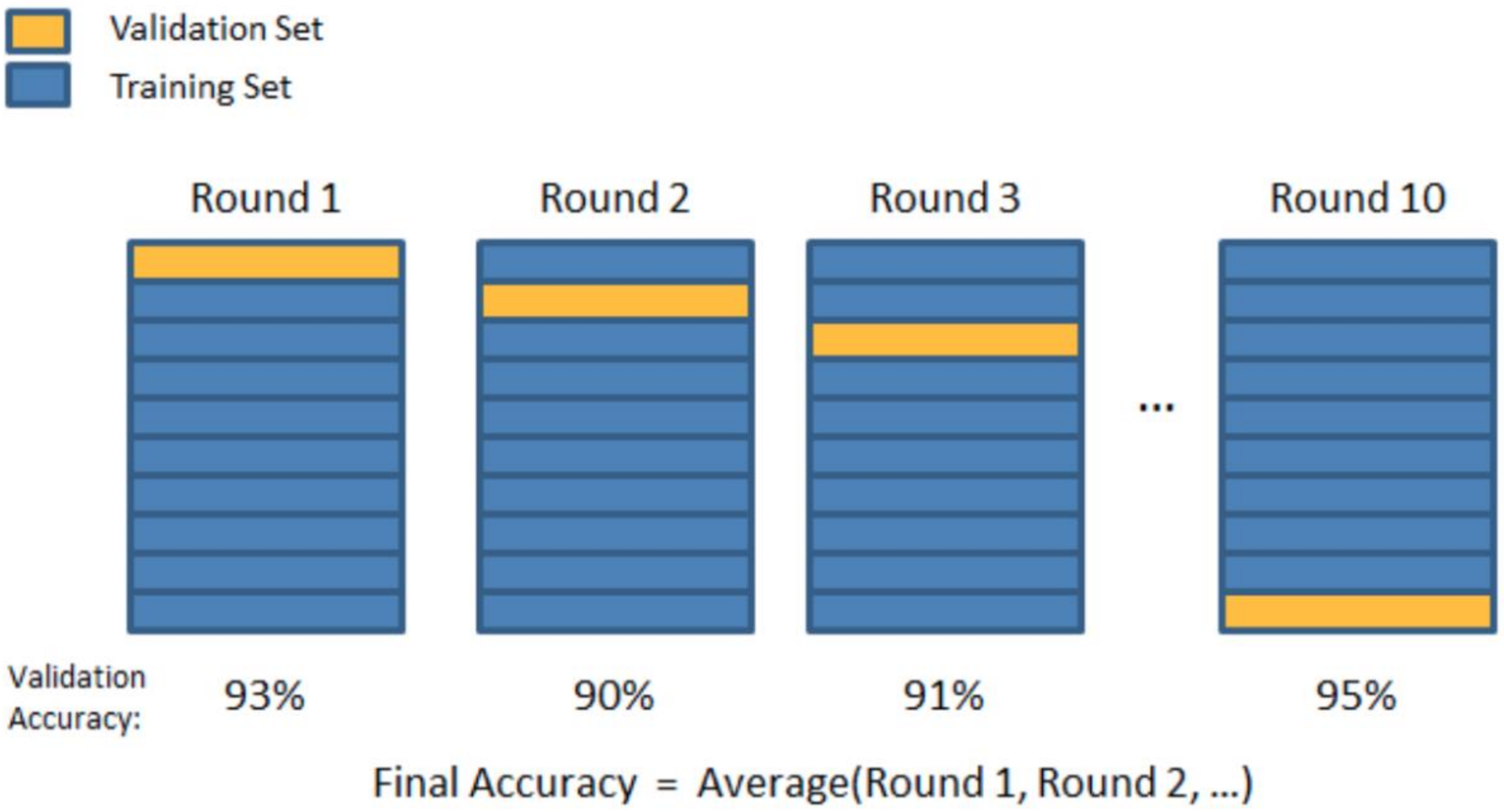
Model Selection

10-Fold Cross-Validation

- Partition the data into 10 folds.
- Use the first fold as the validation set and the remaining folds as the training set.
  - Fit a model by minimizing training set error.
  - Make predictions on the validation set.
- Repeat 10 times with a different fold out each time.
- The average *out-of-fold* error is an estimate of the generalization error.

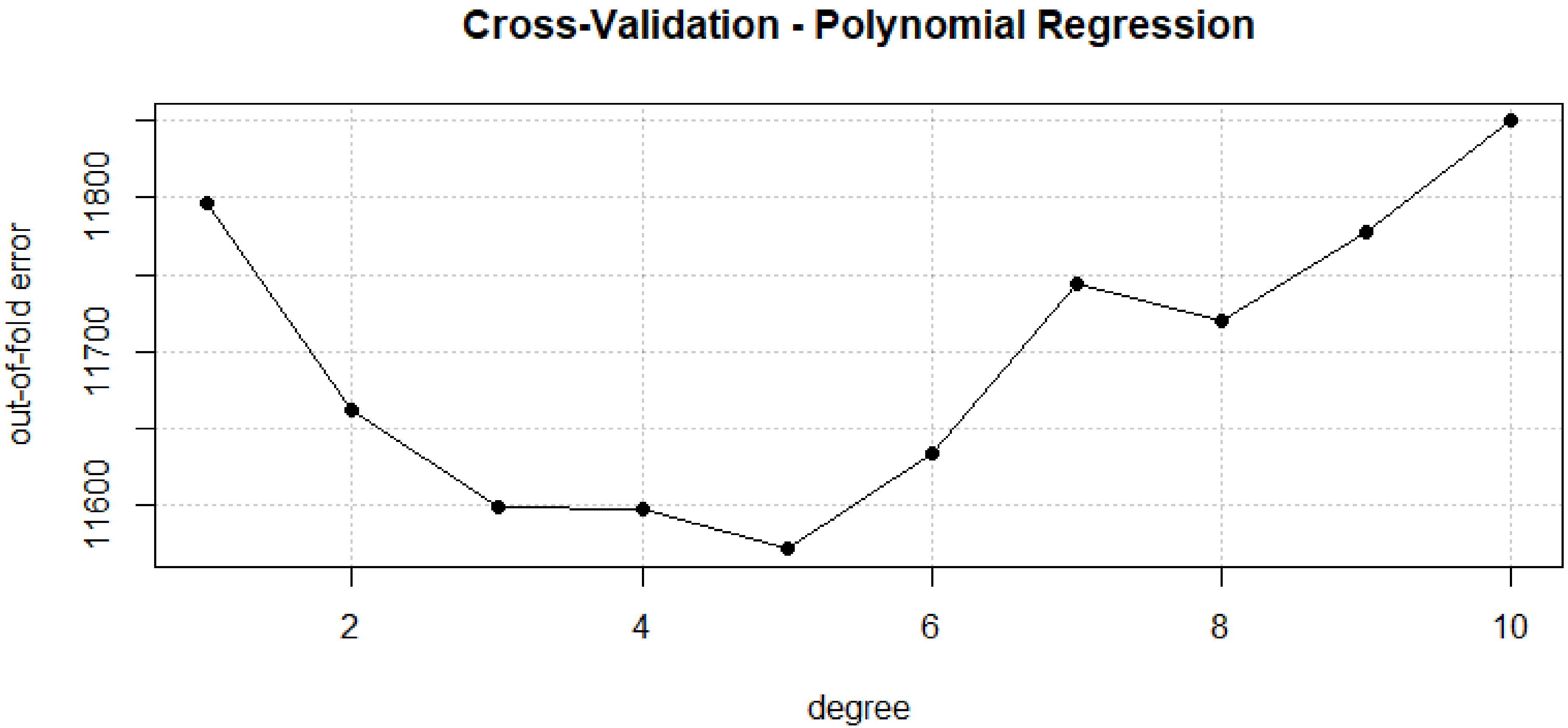
Model Selection

10-Fold Cross-Validation



Model Selection

10-Fold Cross-Validation



Model Selection

Bootstrap aggregating (Bagging)

- Use *bootstrap sampling* (sampling with replacement) to create a training set. All observations not in the training set go in the validation set.
  - Fit a model by minimizing training set error.
  - Make predictions on the validation set.
- Repeat multiple times.
- The average *out-of-bag* error is an estimate of the generalization error.

Model Selection

Bootstrap aggregating (Bagging)

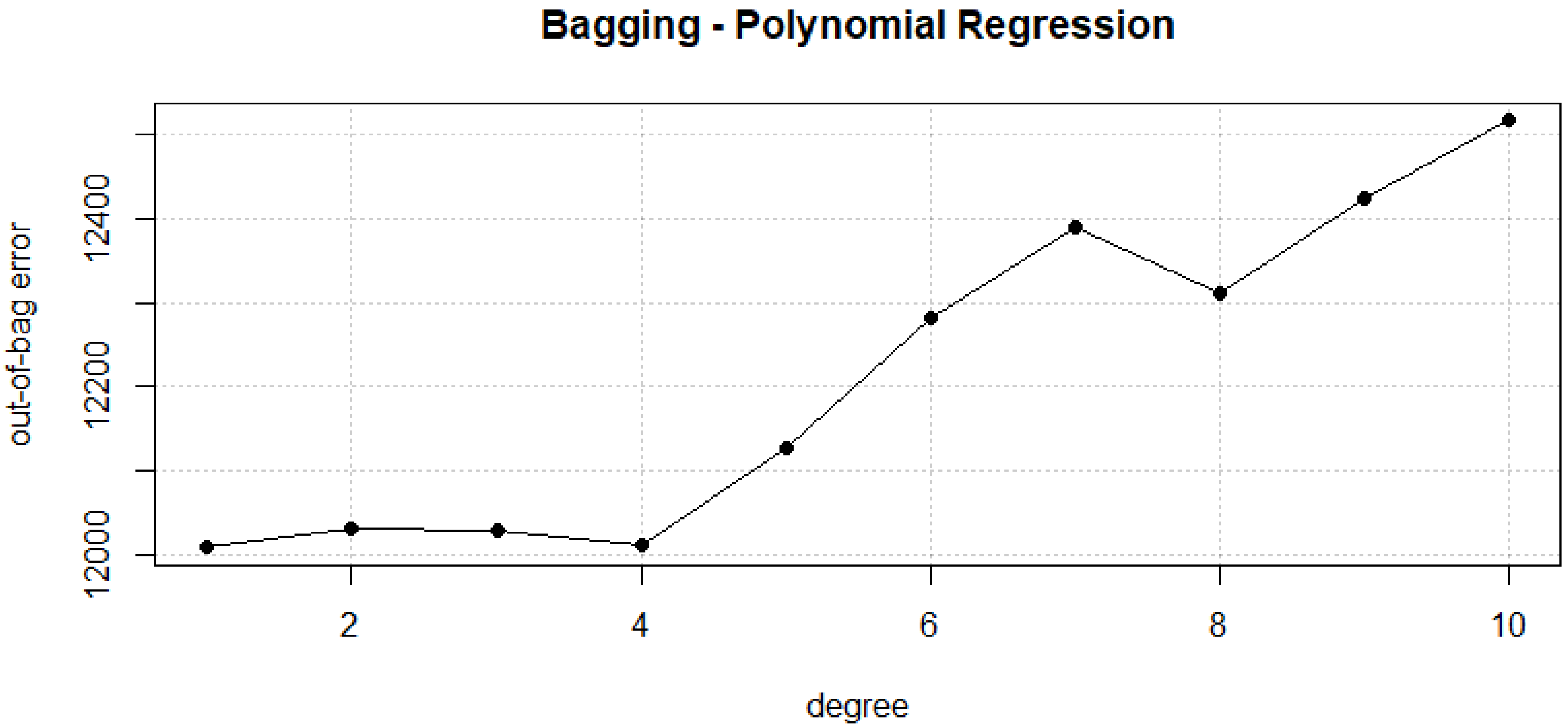


Table of Contents

- Recap
- Basics of R
- Exploratory Data Analysis
- Feature Engineering
- Model Selection
- Resources
- Hands-On Practice

Resources

- **Online Courses**
  Statistical Learning – by Stanford Online (Trevor Hastie, Rob Tibshirani)
  The Analytics Edge – by MITx (Dimitris Bertsimas)
  Machine Learning A-Z – by SuperDataScience Team
- **Free Online Books**
  R for Data Science – by Garrett Grolemund and Hadley Wickham
  Advanced R – by Hadley Wickham
- **Cheat Sheets**
  data.table – by DataCamp
  Miscellaneous – by R Studio

Table of Contents

- Recap
- Basics of R
- Exploratory Data Analysis
- Feature Engineering
- Model Selection
- Resources
- Hands-On Practice

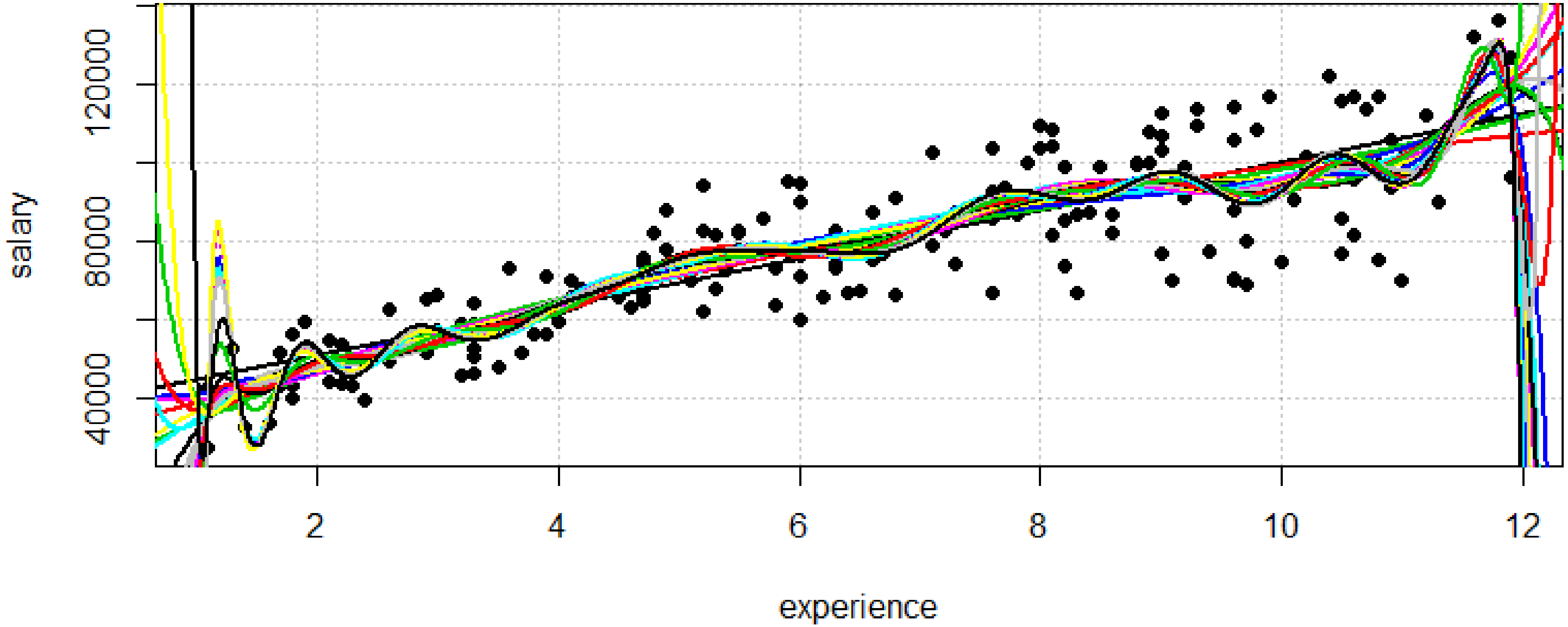
Hands-On Practice

Code - https://github.com/hawaiimachinelearning/into-to-machine-learning-in-r

- **Exploratory Data Analysis** – fit and plot all 1-25 degree polynomials
- **Feature Engineering** – create feature to capture seasonal trend
- **Model Selection** – use cross-validation to tune the `mtry` hyperparameter of the `randomForest` function
- **Model Selection** – use bagging to tune the `alpha` hyperparameter of the `glmnet` function

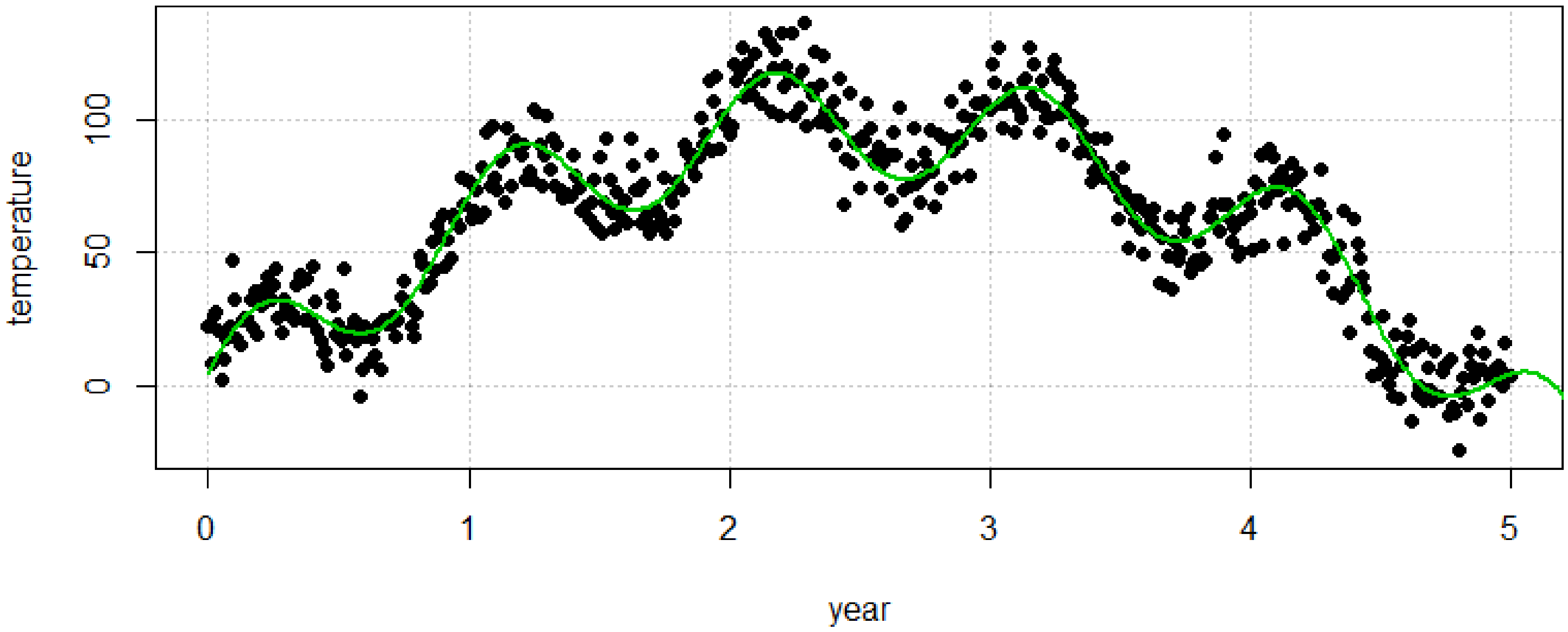
Hands-On Practice

Exploratory Data Analysis

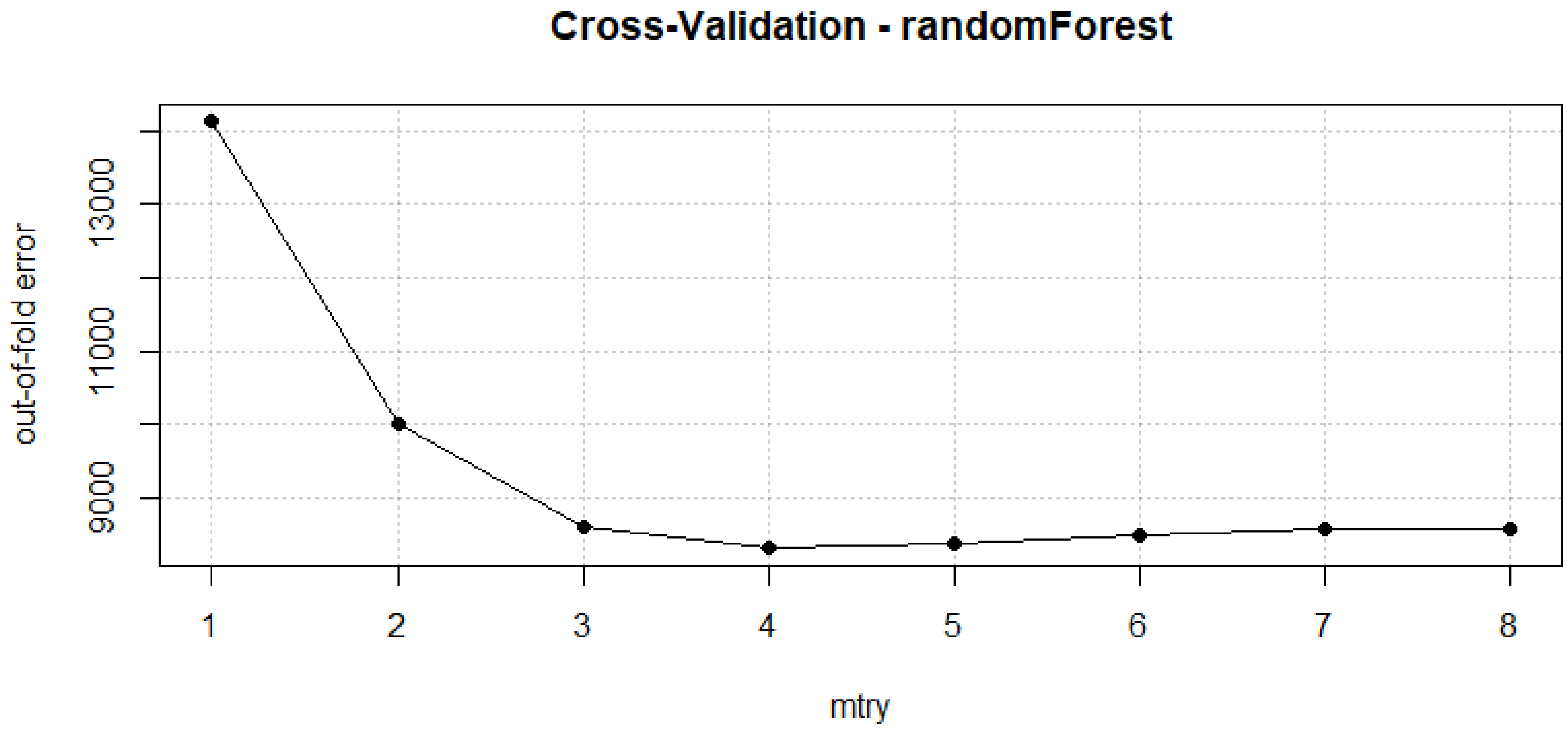


Hands-On Practice

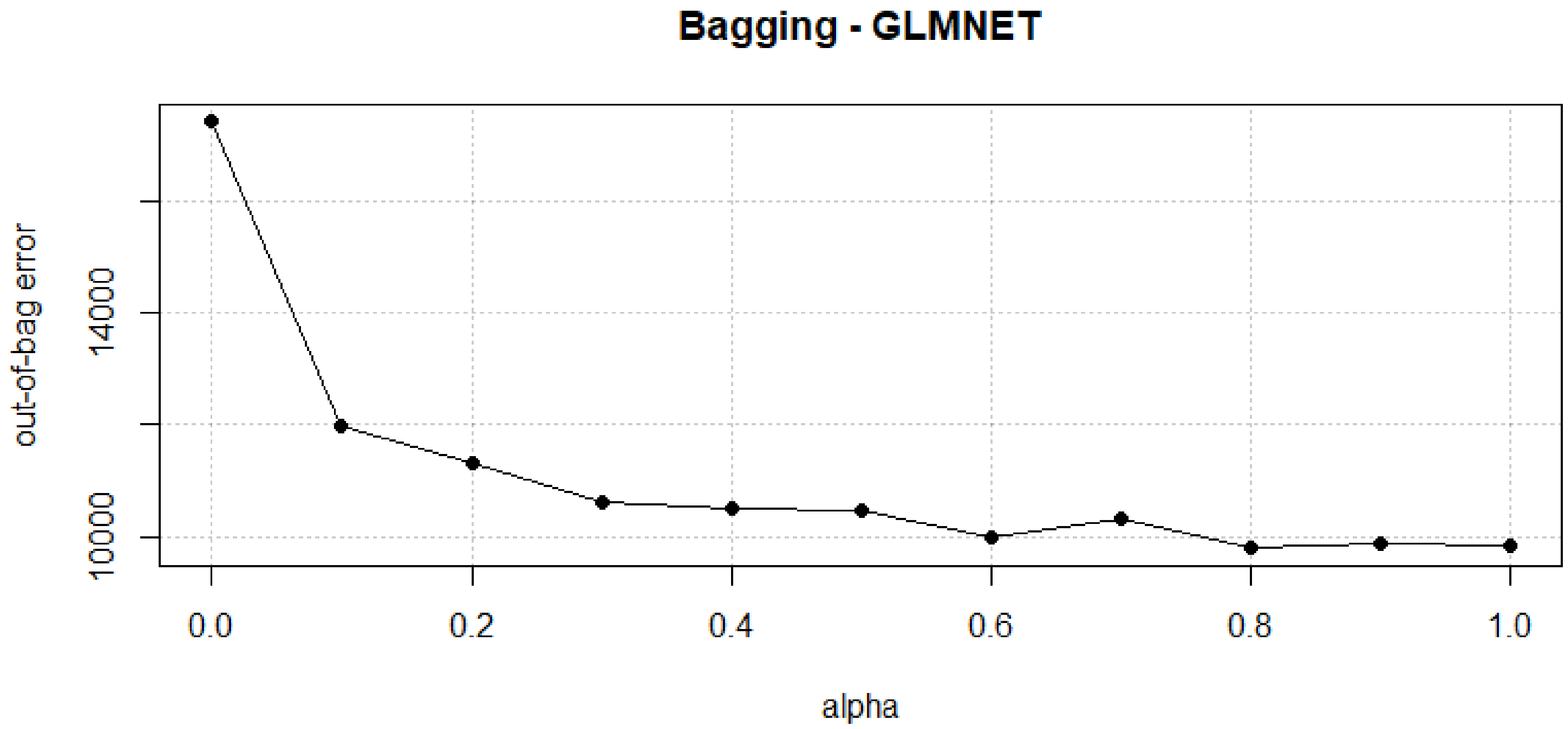
Feature Engineering



Model Selection



Model Selection


Bagging - GLMNET
out-of-bag error
alpha