

# Table of Contents

**About the Author .....xxi**

**About the Technical Reviewer .....xxiii**

**Acknowledgments .....xxv**

**Introduction .....xxvii**

  

**Part I: Getting Started with Google Cloud Platform ..... 1**

**Chapter 1: What Is Cloud Computing? ..... 3**

    Categories of Cloud Solutions ..... 4

    Cloud Computing Models ..... 5

**Chapter 2: An Overview of Google Cloud Platform Services..... 7**

    Cloud Compute..... 7

    Cloud Storage..... 8

    Big Data and Analytics ..... 9

    Cloud Artificial Intelligence (AI) ..... 10

**Chapter 3: The Google Cloud SDK and Web CLI ..... 11**

    Setting Up an Account on Google Cloud Platform ..... 12

    GCP Resources: Projects ..... 14

    Accessing Cloud Platform Services ..... 16

    Account Users and Permissions ..... 16

    The Cloud Shell ..... 17

    Google Cloud SDK ..... 19

TABLE OF CONTENTS

**Chapter 4: Google Cloud Storage (GCS) ..... 25**

    Create a Bucket..... 25

    Uploading Data to a Bucket..... 27

    Delete Objects from a Bucket ..... 30

    Free Up Storage Resource ..... 30

    Working with GCS from the Command Line..... 32

**Chapter 5: Google Compute Engine (GCE) ..... 35**

    Provisioning a VM Instance ..... 35

    Connecting to the VM Instance ..... 41

    Tearing Down the Instance ..... 44

    Working with GCE from the Command Line ..... 45

**Chapter 6: JupyterLab Notebooks ..... 49**

    Provisioning a Notebook Instance..... 49

    Shut Down/Delete a Notebook Instance ..... 53

    Starting a Notebook Instance from the Command Line ..... 54

**Chapter 7: Google Colaboratory ..... 59**

    Starting Out with Colab ..... 59

    Change Runtime Settings ..... 61

    Storing Notebooks ..... 62

    Uploading Notebooks ..... 64

**Part II: Programming Foundations for Data Science ..... 65**

**Chapter 8: What Is Data Science? ..... 67**

    The Challenge of Big Data..... 67

    The Data Science Opportunity..... 68

    The Data Science Process ..... 69

<b>Chapter 9: Python .....</b>	<b>71</b>
Data and Operations .....	71
Data Types.....	72
More on Lists.....	74
Strings .....	77
Arithmetic and Boolean Operations .....	78
Arithmetic Operations.....	78
Boolean Operations .....	78
The print() Statement.....	79
Using the Formatter.....	80
Control Structures.....	80
The if/elif (else-if) Statements.....	80
The while Loop .....	82
The for Loop .....	83
List Comprehensions .....	84
The break and continue Statements .....	85
Functions .....	86
User-Defined Functions .....	87
Lambda Expressions .....	88
Packages and Modules.....	88
import Statement.....	88
from Statement .....	89
<b>Chapter 10: NumPy.....</b>	<b>91</b>
NumPy 1-D Array.....	91
NumPy Datatypes.....	93
Indexing + Fancy Indexing (1-D).....	94
Boolean Mask.....	94
Integer Mask.....	95
Slicing a 1-D Array.....	95

TABLE OF CONTENTS

Basic Math Operations on Arrays: Universal Functions.....	95
Higher-Dimensional Arrays .....	96
Creating 2-D Arrays (Matrices) .....	97
Creating 3-D Arrays .....	98
Indexing/Slicing of Matrices.....	99
Matrix Operations: Linear Algebra.....	100
Matrix Multiplication (Dot Product).....	100
Element-Wise Operations .....	102
Scalar Operation .....	103
Matrix Transposition .....	105
The Inverse of a Matrix.....	106
Reshaping .....	107
Reshape vs. Resize Method.....	107
Stacking Arrays .....	108
Broadcasting .....	110
Loading Data.....	113
<b>Chapter 11: Pandas .....</b>	<b>115</b>
Pandas Data Structures .....	115
Series .....	115
DataFrames .....	117
Data Indexing (Selection/Subsets) .....	120
Selecting a Column from a DataFrame.....	121
Selecting a Row from a DataFrame.....	122
Selecting Multiple Rows and Columns from a DataFrame .....	123
Slice Cells by Row and Column from a DataFrame .....	124
DataFrame Manipulation.....	125
Removing a Row/Column .....	125
Adding a Row/Column .....	127
Data Alignment .....	129
Combining Datasets .....	131

Handling Missing Data .....	132
Identifying Missing Data .....	132
Removing Missing Data .....	133
Imputing Values into Missing Data .....	135
Data Aggregation (Grouping) .....	136
Statistical Summaries .....	138
Correlation .....	139
Skewness .....	139
Importing Data .....	140
Timeseries with Pandas .....	140
Importing a Dataset with a DateTime Column .....	140
Selection Using DatetimeIndex .....	142
Subset Data Columns and Find Summaries .....	144
Resampling Datetime Objects .....	145
Convert to Datetime Datatype Using 'to_datetime' .....	146
The shift() Method .....	147
Rolling Windows .....	148
<b>Chapter 12: Matplotlib and Seaborn .....</b>	<b>151</b>
Matplotlib and Seaborn .....	151
Pandas Plotting Methods .....	151
Univariate Plots .....	152
Line Plot .....	152
Bar Plot .....	154
Histogram/Density Plots .....	155
Box and Whisker Plots .....	157
Multivariate Plots .....	158
Scatter Plot .....	158
Pairwise Scatter Plot .....	160
Correlation Matrix Plots .....	162
Images .....	164

**Part III: Introducing Machine Learning..... 167**

**Chapter 13: What Is Machine Learning?..... 169**

    The Role of Data..... 170

    The Cost of Data..... 170

**Chapter 14: Principles of Learning ..... 171**

    Supervised Learning ..... 171

        Regression vs. Classification..... 172

        How Do We Know that Learning Has Occurred?..... 175

        Training, Test, and Validation Datasets ..... 176

        Bias vs. Variance Trade-Off..... 177

        Evaluating Model Quality ..... 180

        Resampling Techniques..... 191

        Improving Model Performance ..... 195

    Unsupervised Learning ..... 196

    Reinforcement Learning..... 197

**Chapter 15: Batch vs. Online Learning ..... 199**

    Batch Learning..... 199

    Online Learning..... 200

**Chapter 16: Optimization for Machine Learning: Gradient Descent ..... 203**

    The Learning Rate of Gradient Descent Algorithm ..... 204

    Classes of Gradient Descent Algorithm..... 205

    Optimizing Gradient Descent with Feature Scaling..... 205

**Chapter 17: Learning Algorithms..... 209**

    Classes of Supervised Algorithms..... 209

    Unsupervised Algorithms ..... 211

<b>Part IV: Machine Learning in Practice .....</b>	<b>213</b>
<b>Chapter 18: Introduction to Scikit-learn.....</b>	<b>215</b>
Loading Sample Datasets from Scikit-learn .....	215
Splitting the Dataset into Training and Test Sets.....	217
Preprocessing the Data for Model Fitting .....	217
Data Rescaling.....	218
Standardization .....	219
Normalization .....	221
Binarization .....	222
Encoding Categorical Variables .....	223
Input Missing Data.....	227
Generating Higher-Order Polynomial Features .....	228
Machine Learning Algorithms .....	229
<b>Chapter 19: Linear Regression .....</b>	<b>231</b>
The Regression Model .....	231
A Visual Representation of Linear Regression.....	233
Finding the Regression Line – How Do We Optimize the Parameters of the Linear Model? .....	234
How Do We Interpret the Linear Regression Model? .....	234
Linear Regression with Scikit-learn.....	235
Adapting to Non-linearity .....	237
Higher-Order Linear Regression with Scikit-learn .....	238
Improving the Performance of a Linear Regression Model.....	240
<b>Chapter 20: Logistic Regression.....</b>	<b>243</b>
Why Logistic Regression? .....	243
Introducing the Logit or Sigmoid Model.....	245
Training the Logistic Regression Model .....	246
Multi-class Classification/Multinomial Logistic Regression.....	247
Logistic Regression with Scikit-learn .....	248
Optimizing the Logistic Regression Model.....	250

TABLE OF CONTENTS

**Chapter 21: Regularization for Linear Models..... 251**

    How Does Regularization Work ..... 251

    Effects of Regularization on Bias vs. Variance ..... 251

    Applying Regularization to Models with Scikit-learn ..... 252

        Linear Regression with Regularization ..... 252

        Logistic Regression with Regularization ..... 253

**Chapter 22: Support Vector Machines..... 255**

    What Is a Hyperplane? ..... 255

        Finding the Optimal Hyperplane ..... 256

    The Support Vector Classifier ..... 257

        The C Parameter ..... 259

    Multi-class Classification ..... 260

        One-vs.-One (OVO)..... 260

        One-vs.-All (OVA) ..... 261

    The Kernel Trick: Fitting Non-linear Decision Boundaries ..... 262

        Adding Polynomial Features ..... 263

        Kernels ..... 264

**Chapter 23: Ensemble Methods..... 269**

    Decision Trees ..... 269

        On Regression and Classification with CART ..... 270

        Growing a Regression Tree ..... 271

        Growing a Classification Tree ..... 271

        Tree Pruning ..... 272

        Strengths and Weaknesses of CART ..... 272

        CART with Scikit-learn..... 272

    Random Forests ..... 274

        Making Predictions with Random Forests ..... 275

        Random Forests with Scikit-learn ..... 277

    Stochastic Gradient Boosting (SGB) ..... 279

        Tree Depth/Number of Trees ..... 281



Shrinkage .....	281
Stochastic Gradient Boosting with Scikit-learn.....	281
XGBoost (Extreme Gradient Boosting).....	284
XGBoost with Scikit-learn.....	284
<b>Chapter 24: More Supervised Machine Learning Techniques with Scikit-learn ...</b>	<b>287</b>
Feature Engineering.....	287
Statistical Tests to Select the Best $k$ Features Using the SelectKBest Module .....	288
Recursive Feature Elimination (RFE) .....	289
Feature Importances .....	290
Resampling Methods .....	291
k-Fold Cross-Validation .....	291
Leave-One-Out Cross-Validation (LOOCV).....	292
Model Evaluation.....	293
Regression Evaluation Metrics .....	294
Classification Evaluation Metrics.....	297
Pipelines: Streamlining Machine Learning Workflows .....	299
Pipelines Using make_pipeline .....	301
Pipelines Using FeatureUnion.....	302
Model Tuning.....	304
Grid Search.....	304
Randomized Search.....	306
<b>Chapter 25: Clustering.....</b>	<b>309</b>
K-Means Clustering .....	310
Considerations for Selecting $K$ .....	311
Considerations for Assigning the Initial $K$ Points .....	311
K-Means Clustering with Scikit-learn.....	312
Hierarchical Clustering.....	313
How Are Clusters Formed.....	314
Hierarchical Clustering with the SciPy Package .....	317

TABLE OF CONTENTS

**Chapter 26: Principal Component Analysis (PCA) ..... 319**

    How Are Principal Components Computed ..... 319

    Dimensionality Reduction with PCA ..... 321

    Key Considerations for Performing PCA..... 322

    PCA with Scikit-learn ..... 323

**Part V: Introducing Deep Learning..... 325**

**Chapter 27: What Is Deep Learning? ..... 327**

    The Representation Challenge ..... 327

    Inspiration from the Brain ..... 328

**Chapter 28: Neural Network Foundations..... 331**

    The Architecture..... 331

**Chapter 29: Training a Neural Network ..... 333**

    Cost Function or Loss Function..... 336

    One-Hot Encoding ..... 336

    The Backpropagation Algorithm..... 337

    Activation Functions..... 338

        Sigmoid ..... 340

        Hyperbolic Tangent (tanh)..... 341

        Rectified Linear Unit (ReLU)..... 342

        Leaky ReLU..... 342

        Maxout..... 343

**Part VI: Deep Learning in Practice..... 345**

**Chapter 30: TensorFlow 2.0 and Keras ..... 347**

    Navigating Through the TensorFlow API..... 347

        The Low-Level TensorFlow APIs ..... 348

        The Mid-Level TensorFlow APIs ..... 348

        The High-Level TensorFlow APIs..... 352

The Anatomy of a Keras Program .....	355
TensorBoard .....	356
Features in TensorFlow 2.0 .....	358
A Simple TensorFlow Program .....	358
Building Efficient Input Pipelines with the Dataset API .....	359
Linear Regression with TensorFlow .....	361
Classification with TensorFlow .....	365
Visualizing with TensorBoard .....	369
Running TensorFlow with GPUs .....	374
TensorFlow High-Level APIs: Using Estimators .....	381
Neural Networks with Keras .....	383
Using the Keras Sequential API .....	384
Using the Keras Functional API .....	386
Model Visualization with Keras .....	389
TensorBoard with Keras .....	393
Checkpointing to Select Best Models .....	396
<b>Chapter 31: The Multilayer Perceptron (MLP) .....</b>	<b>401</b>
The Concept of Hierarchies .....	401
Choosing the Number of Hidden Layers: Bias/Variance Trade-Off .....	402
Multilayer Perceptron (MLP) with Keras .....	403
<b>Chapter 32: Other Considerations for Training the Network.....</b>	<b>407</b>
Weight Initialization.....	407
Batch Normalization.....	408
Gradient Clipping.....	410
<b>Chapter 33: More on Optimization Techniques .....</b>	<b>411</b>
Momentum.....	411
Variable Learning Rates .....	412
Adaptive Learning Rates .....	413

TABLE OF CONTENTS

**Chapter 34: Regularization for Deep Learning ..... 415**

Dropout ..... 415

Data Augmentation ..... 417

Noise Injection ..... 417

Early Stopping..... 418

**Chapter 35: Convolutional Neural Networks (CNN)..... 423**

Local Receptive Fields of the Visual Cortex ..... 425

Advantages of CNN over MLP ..... 426

    The Convolutional Layer ..... 427

    The Pooling Layer ..... 433

    The Fully Connected Network Layer..... 435

An Example CNN Architecture..... 436

CNN for Image Recognition with TensorFlow 2.0..... 437

**Chapter 36: Recurrent Neural Networks (RNNs)..... 443**

The Recurrent Neuron..... 443

Unfolding the Recurrent Computational Graph ..... 444

Basic Recurrent Neural Network..... 446

Recurrent Connection Schemes..... 448

Sequence Mappings ..... 450

Training the Recurrent Network: Backpropagation Through Time ..... 453

The Long Short-Term Memory (LSTM) Network..... 454

Peephole Connection ..... 456

Gated Recurrent Unit (GRU)..... 457

Recurrent Neural Networks Applied to Sequence Problems..... 458

    Long-Term Recurrent Convolutional Network (LRCN)..... 459

    Encoder-Decoder LSTMs ..... 460

    Bidirectional Recurrent Neural Networks ..... 461

RNN with TensorFlow 2.0: Univariate Timeseries..... 462

RNN with TensorFlow 2.0: Multivariate Timeseries..... 468

<b>Chapter 37: Autoencoders .....</b>	<b>475</b>
Stacked Autoencoders .....	476
Stacked Autoencoders with TensorFlow 2.0 .....	477
Denoising Autoencoders .....	481
 <b>Part VII: Advanced Analytics/Machine Learning on Google Cloud Platform .....</b>	 <b>483</b>
<b>Chapter 38: Google BigQuery .....</b>	<b>485</b>
What BigQuery Is Not .....	486
Getting Started with BigQuery .....	486
Public Datasets .....	489
Running Your First Query .....	490
Loading Data into BigQuery .....	491
Staging the Data in GCS .....	491
Loading Data Using the BigQuery Web UI .....	492
The bq Command-Line Utility .....	496
Loading Data Using the Command-Line bq Utility .....	497
BigQuery SQL .....	499
Filtering .....	499
Aggregation .....	501
Joins .....	502
Subselect .....	504
The Case Against Running Select * .....	505
Using BigQuery with Notebooks on AI Cloud Instance and Google Colab .....	507
BigQueryML .....	509
 <b>Chapter 39: Google Cloud Dataprep .....</b>	 <b>519</b>
Getting Started with Cloud Dataprep .....	519
Using Flows to Transform Data .....	522

TABLE OF CONTENTS

**Chapter 40: Google Cloud Dataflow ..... 537**

    Beam Programming ..... 537

    Building a Simple Data Processing Pipeline ..... 539

**Chapter 41: Google Cloud Machine Learning Engine (Cloud MLE) ..... 545**

    The Cloud MLE Train/Deploy Process ..... 545

    Preparing for Training and Serving on Cloud MLE..... 547

    Packaging the Code for Training on Cloud MLE ..... 548

    The TensorFlow Model ..... 549

    The Application Logic..... 553

    Training on Cloud MLE ..... 558

        Running a Single Instance Training Job ..... 558

        Running a Distributed Training Job ..... 560

        Running a Distributed Training Job with Hyper-parameter Tuning ..... 561

        hptuning\_config.yaml File ..... 562

    Execute Training Job with Hyper-parameter Tuning ..... 563

    Making Predictions on Cloud MLE ..... 565

    Run Batch Prediction ..... 566

    Training with GPUs on Cloud MLE ..... 569

    Scikit-learn on Cloud MLE..... 572

    Move the Data Files to GCS..... 572

    Prepare the Training Scripts..... 573

    Execute a Scikit-learn Training Job on Cloud MLE..... 575

    Create a Scikit-learn Prediction Service on Cloud MLE ..... 577

    Make Online Predictions from the Scikit-learn Model ..... 578

**Chapter 42: Google AutoML: Cloud Vision ..... 581**

    Enable AutoML Cloud Vision on GCP ..... 581

    Preparing the Training Dataset..... 586

    Building Custom Image Models on Cloud AutoML Vision ..... 588

<b>Chapter 43: Google AutoML: Cloud Natural Language Processing .....</b>	<b>599</b>
Enable AutoML NLP on GCP .....	599
Preparing the Training Dataset.....	602
Building a Custom Language Classification Model on Cloud AutoML NLP .....	605
<b>Chapter 44: Model to Predict the Critical Temperature of Superconductors .....</b>	<b>613</b>
The Modeling Architecture on GCP.....	614
Stage Raw Data in GCS .....	615
Load Data into BigQuery for Analytics.....	615
Exploratory Data Analysis .....	617
Spot Checking Machine Learning Algorithms .....	621
Dataflow and TensorFlow Transform for Large-Scale Data Processing .....	624
Training on Cloud MLE .....	636
Deploy Trained Model.....	649
Batch Prediction.....	650
<b>Part VIII: Productionalizing Machine Learning Solutions on GCP .....</b>	<b>653</b>
<b>Chapter 45: Containers and Google Kubernetes Engine .....</b>	<b>655</b>
Docker.....	656
Virtual Machines vs. Containers.....	657
Working with Docker.....	659
Build and Run a Simple Docker Container .....	661
Build the Image .....	661
Run the Container.....	662
Important Docker Commands.....	663
Kubernetes.....	664
Features of Kubernetes .....	665
Components of Kubernetes .....	665
Writing a Kubernetes Deployment File .....	667
Deploying Kubernetes on Google Kubernetes Engine .....	668

TABLE OF CONTENTS

**Chapter 46: Kubeflow and Kubeflow Pipelines..... 671**

    The Efficiency Challenge ..... 672

    Kubeflow ..... 673

        Working with Kubeflow ..... 675

    Kubeflow Pipelines – Kubeflow for Poets ..... 681

        Components of Kubeflow Pipelines ..... 682

        Executing a Sample Pipeline ..... 683

**Chapter 47: Deploying an End-to-End Machine Learning Solution on  
Kubeflow Pipelines ..... 687**

    Overview of a Simple End-to-End Solution Pipeline ..... 688

    Create a Container Image for Each Component ..... 688

    Build Containers Before Uploading to Kubeflow Pipelines..... 689

    Compile the Pipeline Using the Kubeflow Pipelines DSL Language..... 689

    Upload and Execute the Pipeline to Kubeflow Pipelines..... 690

**Index..... 697**