

Indice

Introduzione, di *Emanuele Borgonovo* IX

Parte I **Il metodo**

| | | |
|----------|---|----|
| 1 | Big Data e data lake: le tecnologie di ingestion, storage e calcolo , di <i>Alessandro Rezzani</i> | 3 |
| 1.1 | Introduzione | 3 |
| 1.2 | Panorama tecnologico | 4 |
| 1.2.1 | Il ciclo di vita dei Big Data | 4 |
| 1.2.2 | Strumenti di Data Ingestion | 5 |
| 1.2.3 | Strumenti di storage e organizzazione dei dati | 5 |
| 1.2.4 | Strumenti di trasformazione e analisi dati | 6 |
| 1.3 | L'ecosistema Hadoop | 7 |
| 1.3.1 | Cos'è Hadoop | 7 |
| 1.3.2 | Il file system distribuito | 9 |
| 1.3.3 | I motori di calcolo distribuiti di Hadoop | 10 |
| 1.3.4 | Gli strumenti di data ingestion di Hadoop | 13 |
| 1.3.5 | Hadoop 3 | 14 |
| 1.4 | Spark | 15 |
| 1.4.1 | Cos'è Spark | 15 |
| 1.4.2 | Le componenti di Spark | 15 |
| 1.5 | Altri strumenti | 18 |
| 1.5.1 | Apache Drill | 18 |
| 1.5.2 | Elasticsearch & Kibana | 18 |
| 1.5.3 | Graphana | 19 |
| 1.6 | Il data lake | 21 |
| 1.6.1 | Introduzione | 21 |
| 1.6.2 | Le fasi dell'adozione del data lake in azienda | 22 |

| | | |
|----------|---|----|
| 1.6.3 | L'architettura | 23 |
| 1.6.4 | Lambda architecture e kappa architecture | 25 |
| 1.6.5 | Fattori di successo | 27 |
| 2 | La preparazione dei dati , di <i>Alessandro Recla</i> | 31 |
| 2.1 | La rilevanza della preparazione dei dati per gli advanced analytics | 31 |
| 2.2 | Le fasi della data preparation | 31 |
| 2.3 | Identificazione e classificazione delle variabili | 32 |
| 2.4 | Analisi esplorativa | 35 |
| 2.5 | Trattamento dei missing values | 40 |
| 2.6 | Identificazione e trattamento degli outlier | 44 |
| 2.7 | Trasformazione delle variabili | 46 |
| 3 | Valutazione dei modelli predittivi , di <i>Luca Molteni</i> | 49 |
| 3.1 | I principali modelli di classificazione e di previsione | 49 |
| 3.2 | L'ottimizzazione della stima nei modelli basati sugli alberi di classificazione e di regressione | 51 |
| 3.3 | La valutazione della qualità degli algoritmi di classificazione | 55 |
| 3.4 | La valutazione di qualità predittiva nel modello di regressione logistica | 61 |
| 3.5 | La valutazione di qualità predittiva dei modelli con variabile target numerica | 62 |
| 4 | Metodi di classificazione e regressione ad albero: algoritmi tradizionali ed evoluzioni recenti , di <i>Luca Molteni</i> | 65 |
| 4.1 | Introduzione | 65 |
| 4.2 | Le caratteristiche di base degli algoritmi ad albero | 67 |
| 4.3 | I differenti algoritmi disponibili | 70 |
| 4.4 | Evoluzioni recenti: Random Forest e Gradient Boosting | 72 |
| 5 | Regressione lineare e logistica , di <i>Alessandro Recla</i> | 77 |
| 5.1 | Introduzione ai modelli di regressione lineare e logistica | 77 |
| 5.2 | Il modello di regressione lineare: definizione del modello | 78 |
| 5.3 | Il modello di regressione lineare: stima del modello | 80 |
| 5.4 | Il modello di regressione lineare: valutazione del modello | 81 |
| 5.5 | Il modello di regressione logistica binaria: definizione del modello | 85 |
| 5.6 | Il modello di regressione logistica: stima del modello | 86 |
| 5.7 | Il modello di regressione logistica: valutazione del modello | 89 |
| 6 | Le reti neurali, nascita, diffusione e funzionamento , di <i>Daniele Tonini e Francesco Tuscolano</i> | 93 |
| 6.1 | Introduzione | 93 |

| | | |
|-----|---|-----|
| 6.2 | Il funzionamento delle unità di calcolo | 97 |
| 6.3 | L'introduzione del bias | 104 |
| 6.4 | La back-propagation e il metodo della discesa del gradiente | 105 |
| 6.5 | Overfitting e generalizzazione | 109 |

Parte II

Applicazioni selezionate

| | | |
|----------|---|-----|
| 7 | Modelli di churn prediction: un'applicazione al settore bancario, di <i>Alessandro Recla</i> | 115 |
| 7.1 | Introduzione al problema | 115 |
| 7.1.1 | Applicazione al settore bancario | 116 |
| 7.2 | Preparazione dei dati per l'applicazione dei modelli | 116 |
| 7.2.1 | Applicazione al settore bancario | 118 |
| 7.3 | Stima del modello | 119 |
| 7.3.1 | Applicazione al settore bancario | 120 |
| 7.4 | Messa in produzione del modello | 125 |
| 7.4.1 | Applicazione al settore bancario | 127 |
| 8 | Social media analytics e sentiment analysis: un'applicazione al mercato televisivo, di <i>Luca Molteni</i> | 129 |
| 8.1 | Introduzione | 129 |
| 8.2 | Il caso: previsione dell'ascolto con dati Twitter | 130 |
| 8.3 | La preparazione dei dati per l'analisi | 134 |
| 8.4 | La sentiment analysis | 135 |
| 8.5 | La strutturazione del database per l'analisi predittiva dell'audience | 136 |
| 8.6 | I risultati delle analisi e le implicazioni manageriali | 138 |
| 9 | L'impatto delle immagini sul livello di interesse per un'offerta digital: applicazione al mercato immobiliare, di <i>Luca Molteni</i> | 141 |
| 9.1 | Gli scenari sullo sfondo: evoluzione del settore immobiliare negli ultimi 10 anni e degli advanced analytics nell'era dei Big Data | 141 |
| 9.1.1 | Evoluzione del settore immobiliare negli ultimi 10 anni | 141 |
| 9.1.2 | La rilevanza dell'image analytics nel digital marketing | 143 |
| 9.2 | Il caso applicativo: analisi di alcuni annunci immobiliari di RentHop applicando l'image analytics | 146 |
| 9.2.1 | I dati utilizzati per l'analisi | 146 |
| 9.2.2 | La pre-elaborazione quantitativa delle immagini | 147 |
| 9.3 | I risultati dell'analisi | 149 |
| 9.3.1 | I modelli predittivi del livello di interesse dell'annuncio | 149 |
| 9.4 | Conclusioni | 152 |

| | |
|---|-----|
| 10 L'uso del multilayer perceptron per la prevenzione delle frodi bancarie: un caso applicativo, di <i>Daniele Tonini</i> e <i>Francesco Tuscolano</i> | 153 |
| 10.1 Lo scenario di fondo: la prevenzione dalle frodi nell'era dei Big Data | 153 |
| 10.2 Lo sviluppo di un modello di advanced analytics attraverso KNIME Analytics Platform | 154 |
| Bibliografia | 163 |
| Autori | 167 |