Data Analytics / Big Data — Notes

Overview Summary

Businesses collect data in order to improve their business operations. Some operations include spotting sales trends, developing smarter marketing campaigns, accurately predicting customer behavior, and predicting and protecting against fraud. Ultimately, data mining (data collection) can be used in any context whenever there is an interest in identifying and exploiting predictable outcomes.

Course One: Statistical machine learning techniques to evaluate online customer sales and market research data in order to provide insights into customer buying patters and preferences. This will help determine what products to offer and how to offer them.

Course Two: Use Python to a solve a recurring problem related to a credit provider.

Course Three: Analyze different customer demographics to determine how they impact what brands a customer might choose. Develop a predictive model that can drive product recommendation systems.

Course Four: Advanced analytics and visualization via IoT based on a problem of understanding energy consumption over a four-year period.

Course Five: Sentiment analysis of an extremely large data set using cloud-based computing, machine learning tools and Apache Spark.

“What You’ll Learn”

* Identify types of business problems for which data analysis can provide significant insights in support of business decision-making.
* Translate business objectives into analytical opportunities using data mining.
* Apply statistical machine learning tools and methods to different kinds of data.
* Select and justify appropriate types of data analysis and statistical procedures
* Apply data analytics in eCommerce (e.g., understanding customer behavior, segmenting customers by key demographic factors, selecting new products strategically and predicting their profitability).
* Become broadly competent in the use and evaluation of statistical machine learning techniques of classification, regression and association.
* Apply dimensionality reduction methods to broad datasets to reduce their complexity prior to modelling
* Identify and solve collinearity through feature engineering and feature selection
* Acquire, process, and analyze extremely large data sets using cloud-based data mining methods to discover patterns or do data exploration.
* Interpret the results of data analysis to make predictions and to establish the reliability of those predictions.
* Communicate data mining results to management and other non-technical audiences.
* Construct, justify, and apply custom data science processes
* Apply parametric and non-parametric machine learning models
* Avoid overfitting models
* Develop a data analytics portfolio showcasing work and highlighting skills

Even if you are not directly responsible for data mining, its ever-increasing prevalence in the business world means that you will likely be working with others who are involved in its use. This requires that you be able to speak with them knowledgably about data mining-from theory to practical use and strategy.

The Data Analytics/Big Data Program will equip you with the necessary skills and knowledge to gain entry-level employment as a Data Analyst. You will be prepared to join a team to perform analyses that turn data into business value.

Full Course Description

Course 1: Data Analytics with Python: Understanding Customers

In this course students will be working for Blackwell Electronics as data analysts. The students' job is to use data mining and machine-learning techniques to investigate the patterns in Blackwell's sales data and provide insight into customer buying trends and preferences. The inferences students draw from the patterns in the data will help the business make data-driven decisions about sales and marketing activities.

Students will use Python and various Data Analytics libraries to understand the relationship between customer demographics and purchasing behavior. Students will present their findings to management, explaining their insights and suggestions for data mining process improvements.  
   
Duration: 6 weeks Part-Time / 3 weeks Full-Time  
Prerequisite: none

Course 2: Data Science with Python: Predicting Customer Credit Limits

In this course, you are a Data Scientist for a company called Credit One, which is a third-party credit rating authority that provides retail customer credit approval services to businesses.

Credit One has tasked you with examining current customer demographics to better understand what traits might relate to whether or not a customer is likely to default on their current credit obligations. Understanding this is vital to the success of Credit One because their business model depends on customers paying their debts.  
   
Your job as a Data Scientist will be to identify which customer attributes relate significantly to customer credit limits and to build a predictive model that Customer One can use to better predict the amount of credit certain customers should be assigned, compared to previously implemented models. You will use machine learning regression methods in Python for this task.  
   
   
Duration: 8 weeks  
Prerequisite: Data Analytics and Visualization or equivalent experience

Course 3: Data Analytics with R: Recommendation Systems and Market Basket Analysis

Students will continue to work as data analysts at Blackwell Electronics. Their job is to extend Blackwell's application of data mining methods to develop predictive models.

In this course, students will use the R statistical programming language augmented with machine learning packages to predict different product types a customer will be likely to buy. Next, students will create a recommender system that recommends additional purchases based on customer choices. Finally, students will present to management, explaining their insights and suggestions for data mining process improvements.  
   
Duration: 6 weeks  
Prerequisite: Data Analytics, Classification and/or Regression

Course 4: Data Science with R: Energy Consumption and WiFi-Locationing

Increasingly, technology companies are applying data analytics techniques to the masses of data generated by devices such as smart phones, appliances, vehicles, electric meters, et cetera – the "Internet of Things". The ability to deal with data of these types will prove to be a high-demand skill for data analysts as applications of these sorts proliferate.  
   
In this course, students will be working for an Internet of Things technology start-up that wants to use Data Analytics to solve two difficult problems in the physical world:

* Smart energy usage: Modeling patterns of energy usage by time of day and day of the year in a typical residence whose electrical system is monitored by multiple sub-meters.
* Indoor locationing: Determining a person’s physical position in a multi-building indoor space using wifi fingerprinting

Students will learn to use the R statistical programming language to perform visualizations, then to generate descriptive statistics and predictive models using time series regression techniques and statistical classifiers.  
   
Finally, students will present the results to the start-up’s management, explaining strengths and weaknesses of the approaches that were implemented and making suggestions for further improvement.  
   
Duration: 6 weeks  
Prerequisite: Data Analytics, Machine Learning, Data Visualization

Course 5: Big Data with Spark: Text Mining and Sentiment Analysis

In this course, students will be working as data analysts for Alert Analytics, a data analytics consulting firm. Their client, Helio, developed a suite of smartphone medical apps for use by aid workers in developing countries. The government agency will be providing workers with technical support services, but they need to limit the support to a single model of smart phone and operating system. Helio has created a short list of devices that are all capable of executing the app suite's functions. To narrow this list down to one device, Helio has engaged Alert Analytics to conduct a broad-based web sentiment analysis to gain insight into users' attitudes towards the devices. The students’ job is to conduct this analysis.

Students will use the Amazon Web Services (AWS) Elastic Map Reduce (EMR) platform to run a series of Hadoop streaming jobs and R to develop a predictive model that will infer user sentiment towards devices from the lexical content of web pages extracted from the Common Crawl of the World-Wide Web.

Lastly, students dive into a slightly more complicated project and go further than counting the instances or number of times words appear in various documents, but now must dive into the various constructs of text mining and text context altogether. Fortunately, the business scope and criteria for success for this project is straightforward: can the contents of a tweet be classified as having a certain level of sentiment or not?

Students will perform this work using Apache Spark, which is an analytics engine for large-scale data processing. Since the data they’ll be working with is likely too large to work with locally they’ll be using Microsoft Azure and Databricks, which uses a Spark cluster in the cloud, for the heavy lifting on this project.

Duration: 8 weeks

*Prerequisite:* Data Science with R: Energy Consumption and WiFi-Locationing