Case Study: Lila's Journey to Becoming a Data Scientist: Her Working Approach on the First Task

This case study explores the data scientist's career path and key attributes, highlighting the skills, education, and experiences required to excel in this dynamic field. We'll follow the story of Lila, a fictional individual who aspires to become a successful data scientist.

There will be a quiz after this reading based on the contents of this case study.

Education and Skill Acquisition

With an economics undergraduate degree and a substantial data analysis background, Lila finds data science and its potential to drive meaningful change captivating. Inspired by her experiences, she makes a determined decision to transition her career and step into the role of a data scientist.

Lila realizes that to embark on her data science journey, she needs to enhance her skills and knowledge. She enrolled in the IBM Data Science Professional Certificate online program that covers key topics like statistics, machine learning, data analysis, and programming languages like Python and SQL. She diligently completes coursework and practices her coding skills on real datasets.

Building a Strong Foundation

As she progresses in her studies, Lila gains a deep understanding of data science fundamentals such as data manipulation and visualization with Python libraries like NumPy, Pandas, and Matplotlib. This strong foundation equips her with essential skills for data analysis.

Visualization for Storytelling

Lila knows she must communicate her findings effectively, so she learns which types of data visualizations will be most informative. She learns to create charts and graphs that visually represent data like sales trends, customer segmentation, and product popularity, allowing stakeholders to grasp the data's significance. These visualizations help in storytelling and decision-making.

Hands-On Experience

Lila understands that practical experience is invaluable in data science. She started participating in Kaggle competitions and working on personal data projects. These experiences expose her to real-world data problems and help her develop problem-solving skills. Furthermore, she created her GitHub account and uploaded her projects to build her profile.

Data Wrangling and Preprocessing

Lila learns that data scientists spend a significant portion of their time on data cleaning and preprocessing. She worked on various datasets, learned data preprocessing as she used sed NumPy and pandas Python libraries, and became skilled in handling missing data, outlier detection, and feature engineering to improve model performance.

Communication and Storytelling

Recognizing that data scientists must communicate their findings effectively, Lila honed her data storytelling skills. She learned various tools like matplotlib and plotly while she pursued her IBM Data Science Professional Certificate. She learned how to create compelling visualizations and present her insights in a clear and understandable manner.

Networking and Collaboration

Lila actively participates in data science communities and attends meetups and conferences. She collaborates on open-source projects, connects with fellow data scientists, and gains exposure to various industries when she attended the IBM TechXchange Conference.

Domain Expertise

Understanding that domain knowledge is crucial, Lila chooses a niche that aligns with her interests. She looks deeply into several domains, including e-commerce, healthcare, finance, and several other fields to which she could apply her data science skills effectively. Since her master's in economics, she chose e-commerce as her core domain to land herself a data science career.

Landing the First Job

After months of preparation, Lila started applying for data scientist positions. She tailors her resume to highlight her relevant skills and projects. Her online portfolio showcases her capabilities and demonstrates her commitment to the field.

Lila's Approach to Working on Her First Task as a Data Scientist

As a newly hired junior data scientist at a retail company, Lila uses data insights to improve customer service. Her first assignment involves diving into customer data to identify patterns and anomalies that could impact customer service. She uses data analysis to enhance the overall customer experience.

Dataset Selection and Sourcing

In the initial phase of her data science journey, Lila faced the challenge of selecting a suitable dataset and procuring it from different sources. Apart from the historical data available for the organizations for the past four years, she scoured various repositories, websites, and databases to find the right datasets for her project. Upon collecting data from diverse sources, Lila encountered another crucial decision point. She had to decide how to harmonize and integrate these disparate datasets into a cohesive whole. She reached out to product professionals, data engineers, and domain specialists, seeking their input and expertise in merging datasets.

Data Understanding and Cleaning

Lila begins by importing the dataset into her data analysis environment using Python and SQL. She loads the data and examines the first few rows to understand its structure and contents. Upon acquiring the dataset, Lila encounters her first challenge: data cleaning. Lila checks for missing values, duplicates, and outliers in the dataset. She addresses missing data by imputing or removing rows or columns with missing values. Outliers are identified and treated appropriately based on their impact on the analysis.

Exploratory Data Analysis (EDA)

As she delves into exploratory data analysis, Lila faces numerous choices. She must determine which summary statistics, visualizations, and distribution analyses will best reveal insights into customer behavior and sales trends. Each choice she makes during EDA influences the story the data will tell. Lila conducts EDA to gain insights into the dataset. She generates summary statistics and visualizations (histograms, scatter plots) and explores the distribution of variables. EDA helps her understand customer behavior, popular products, and sales trends.

Feature Engineering

Lila recognizes the potential for feature engineering to enhance her analysis. She assesses whether creating new features, such as calculating total purchase amounts, will improve the dataset's utility for her project.

Statistical Analysis, Machine Learning

Lila evaluates whether statistical tests or machine learning algorithms are necessary. She employs regression analysis to understand relationships between variables and explore machine learning models for demand forecasting or customer segmentation tasks. Lila also performs statistical tests to uncover patterns in the data. She uses regression analysis to understand the impact of unit price on sales.

Presentation and Reporting

At the culmination of her analysis, Lila faces the challenge of presenting her findings. Lila compiles her analysis and findings using a Jupyter Notebook into a comprehensive report and presentation. She highlights actionable insights and recommendations for the e-commerce platform's stakeholders.

Continuous Learning

After completing her first project, Lila continues to refine her skills, explores more complex datasets, and tackles increasingly challenging data science tasks.

Machine Learning Skills

Although Lila took an introductory course on Machine Learning as part of the IBM Data Science Professional Certificate, the field intrigues her, and she wants to develop her skills further by taking the IBM Machine Learning Professional Certificate. She identified Machine Learning Repository datasets in the course and experimented with various algorithms. Lila dives into machine learning to excel as a data scientist, wherein she studies various algorithms, such as linear regression, decision trees, and deep learning models. She continues to gain expertise in selecting and fine-tuning algorithms based on specific data problems.

Skills Network

IBM