A Comprehensive Exploration of the CRISP-DM Methodology: An Analysis on the Data Science Salaries Dataset

Shawn Chumbar (generated with help of ChatGPT)

September 26, 2023

Abstract

The Cross Industry Standard Process for Data Mining (CRISP-DM) has established itself as a leading framework for guiding the data mining processes across various industries. The methodology offers a structured approach to planning and executing data mining projects. In this paper, we apply the CRISP-DM methodology to a dataset detailing data science salaries, aiming to derive insights about the determinants of salary and trends within the profession. The analysis highlights the value of following a systematic process, while simultaneously shedding light on the factors influencing data science remuneration.

1 Introduction

The explosion of data in the modern age has precipitated the need for structured and robust methodologies to process and derive insights from this vast ocean of information. One such methodology that has gained widespread acceptance is the CRISP-DM. This paper provides an overview of the CRISP-DM process and its phases, and employs this methodology to analyze the data science salaries dataset.

2 CRISP-DM Methodology

The CRISP-DM methodology comprises six main phases:

1. **Business Understanding**: This phase involves understanding the problem domain, objectives, and requirements from a business perspective.
2. **Data Understanding**: Involves collecting, describing, and exploring the data.
3. **Data Preparation**: Data is cleaned, transformed, and made ready for modeling.
4. **Modeling**: Various modeling techniques are applied and their parameters are calibrated.
5. **Evaluation**: The models are evaluated in the context of the business objectives.
6. **Deployment**: Insights and models are deployed for business usage.

3 Data Understanding

To begin our analysis using CRISP-DM, we must first understand the dataset provided.

4 Data Preparation

Before proceeding with the analysis, it is essential to ensure that the data is in the right format and that any inconsistencies or missing values are addressed.

5 Modeling and Analysis

Given that the primary objective is to understand the determinants of data science salaries and trends within the profession, we'll perform exploratory data analysis (EDA) to uncover patterns, relationships, and insights.

5.1 Salary Distribution

Salaries tend to increase with higher experience levels. Senior roles command higher median salaries compared to mid-level and junior roles. The role of Product Data Analyst tends to have a lower median salary.

5.2 Salary by Company Size and Location

Employees in large companies tend to have a higher median salary compared to those in medium and small companies. Employees in the US and JP (Japan) appear to have the highest median salaries among the top locations.

6 Evaluation

The CRISP-DM methodology emphasizes evaluating results against business objectives. Experience level, job title, company size, and location play a significant role in determining salaries.

7 Conclusion

The CRISP-DM methodology provided a structured approach to analyze the data science salaries dataset. These findings can help aspiring data scientists, recruiters, and companies in benchmarking and setting expectations around compensation.

Acknowledgments

We thank the creators of the data science salaries dataset for making this information available for research purposes.