

Deep Learning Pipeline Approach



fulltimeSF

Now interactive!

| | | | |
|---|----------------------|---------------|--|
| | | | |
| 1 | Data Scientist - CA | Bind Benefits | <div><p>Positi on |
| 2 | Staff Data Scientist | Course Hero | The Role: |
| | position | company | description |

Do a deep dive on the application of the CRISP-DM methodology as if you were in giving a presentation to students who are in high school using the dataset. The objective is to be able to guide each students through the process in hopes they will grasp the principles and technique. When providing examples, do not use code but explain the process of what you would do.

Starting with data understanding, how would you approach a structured exploratory data analysis (EDA) while educating students on the importance of data visualization, summary statistics, and pattern detection? Demonstrate how to clean and preprocess the data, making informed decisions about handling missing values, outliers, and data transformation.

Next, explain feature selection and how to apply dimensionality reduction methods in this scenario. Include insights into how clustering might reveal hidden patterns, and how you would deal with outliers.

Then, transition to building predictive models, starting with baseline regression models and progressively integrating more complex machine learning techniques, ensuring students understand model comparison and evaluation through proper metrics like RMSE, MAE, or accuracy. Ensure that for each step, you explain how limited computational resources can be accounted for by chunking tasks efficiently.

Throughout this exercise, make sure that each phase aligns with the CRISP-DM process. At the end of each step, summarize where you are in the methodology and clearly outline the next steps, always keeping the overall objective in mind. Provide detailed yet manageable segments for students to absorb each concept thoroughly before moving on.



Presentation of CRISP-DM Process: Data Understanding and Exploratory Data Analysis (EDA)

Introduction to CRISP-DM

The **CRISP-DM (Cross-Industry Standard Process for Data Mining)** methodology is a structured approach to solving data science problems. It consists of six key phases:

1. **Business Understanding**
2. **Data Understanding**
3. **Data Preparation**
4. **Modeling**
5. **Evaluation**
6. **Deployment**

For high school students, I'll focus on simplifying the concepts to make them approachable. Using the provided dataset, which contains information on **Data Scientist positions in San Francisco**, we'll go through each step.

Step 1: Data Understanding

Objective: Before working with data, it's important to understand what kind of information is available. In our dataset, we have columns like:

- ``position``: Job title
- ``company``: The name of the company offering the position
- ``description``: A summary of the job listing
- ``numreview``: The number of reviews for the company (may have missing values)
- ``location``: Where the job is located

Exploratory Data Analysis (EDA) is an essential first step in data understanding. This includes:

- **Visualizing the data:** Simple bar charts or histograms can show how frequently different positions or companies appear.
- **Summary statistics:** Basic statistics like the mean, median, and mode help you get a feel for the numerical data.
- **Detecting patterns:** Are there specific companies that offer more jobs than others? Are job locations clustered in certain areas?

Step 2: Data Cleaning and Preprocessing

During EDA, you might notice missing or unusual data. Let's talk about how to handle that:

- **Missing Values:** In the ``numreview`` column, some companies don't have reviews. There are a few ways to handle this:
 - **Remove rows** with missing data, but this could lead to losing valuable information.
 - **Imputation:** Replace missing values with an average or median. For example, if most companies have around 15 reviews, we might fill missing values with 15.

- **Outliers:** In the ``numreview`` column, we might see some companies with an unusually high or low number of reviews. We have to decide whether these outliers are valid (e.g., a very large company might have many reviews) or if they should be treated as errors.
- **Data Transformation:** If necessary, we could transform categorical data (like company names) into numerical formats using encoding techniques, which help in the modeling phase.

Step 3: Feature Selection and Dimensionality Reduction

Feature Selection helps us identify the most important columns or features in the dataset that will influence the outcome we're interested in. For instance:

- Is ``location`` a good predictor of job reviews or job availability?
- Does the ``description`` provide keywords that might influence how competitive a job position is?

We could apply **dimensionality reduction techniques**, like **PCA (Principal Component Analysis)**, to reduce the number of variables, focusing on the most significant ones. This simplifies the dataset without losing important information.

Step 4: Clustering and Hidden Patterns

Now, imagine we want to find hidden patterns in the dataset, such as:

- Are there clusters of job positions that are similar based on their descriptions or locations?
- We could use **clustering techniques** (like K-means clustering) to group job positions together. For example, we might discover that data science jobs in healthcare are more clustered in specific regions of the city, while finance-related jobs are elsewhere.

This helps reveal trends and relationships that aren't immediately obvious.

Step 5: Building Predictive Models

We could move on to predictive modeling once the data is prepared and cleaned. For simplicity:

- **Baseline Models:** Start with a simple regression model, predicting, for example, the number of reviews a job might have based on its location or position.

You can explain that this type of model tries to find a straight line that fits the data best. The closer the predicted values are to actual values, the better the model.

- **Model Evaluation:** To compare models, we use metrics like:
 - **RMSE (Root Mean Squared Error):** Measures the difference between the predicted and actual values. Lower RMSE means better predictions.
 - **MAE (Mean Absolute Error):** Similar to RMSE but less sensitive to outliers.

As we progress, we can introduce more complex models like **decision trees** or **random forests**, which are better at capturing complex relationships in the data.

Step 6: Model Comparison and Evaluation

It's important to compare different models and choose the best one based on performance. In addition to metrics, we might consider how easy a model is to understand and how well it works with limited data or resources.

Step 7: Finalizing the CRISP-DM Process

At the end of the CRISP-DM process, we evaluate the results to ensure that they meet the initial business goals, such as predicting trends in data scientist jobs or understanding where high-demand roles are located.

Summary of Each Phase

- **Business Understanding:** What's the problem we're solving? In our case, it could be predicting job trends or analyzing job markets.
- **Data Understanding:** We explored the dataset to understand its structure and content.
- **Data Preparation:** Cleaned and transformed the data for analysis.
- **Modeling:** Built and evaluated different models to predict outcomes.
- **Evaluation:** Compared the models and refined the best one.
- **Deployment:** The model is ready to be used for real-world predictions.

This step-by-step breakdown gives students a manageable way to understand the entire process, connecting theory to practical outcomes.



Give me an image on Data scientist in the SF



Here is the image of a data scientist working in San Francisco. Let me know if you'd like any modifications or additional details!

ChatGPT can make mistakes. Check important info.