Case Study: Data Science at SANDAG – Enhancing Transportation Modeling and Data Analysis

Introduction

This case study details my work as a Data Scientist at the San Diego Association of Governments (SANDAG), where I contributed to optimizing and analyzing large-scale transportation models. My role centered around enhancing data pipelines, performing complex data analyses, and helping the organization improve its transportation planning using advanced modeling techniques. During my tenure, I worked on comparing outputs from different transportation simulation tools, improving data quality control mechanisms, and enabling more data-driven decision-making.

Key Responsibilities

As a Data Scientist at SANDAG, my responsibilities spanned across several critical areas:

- **Data Modeling & Analysis:** Designed and analyzed complex transportation models, focusing on trip calculations and ridership predictions based on historical and simulated data.
- **Model Comparison & Reporting:** Led initiatives to compare outputs from the EMME model (used for transportation forecasting) with new simulation models (OpenPaths 2024), providing insights into the accuracy and performance of the models.
- **Data Quality Control:** Developed automated systems for detecting data anomalies, specifically focusing on issues like negative or zero values in important transit columns (such as `BASEIVTT`), which could affect the quality and accuracy of transportation forecasts.
- **Visualization & Reporting:** Transformed raw data into insightful visualizations and reports to communicate key findings to stakeholders, enabling informed decision-making in transportation planning.
- **Collaboration & Communication:** Collaborated with cross-functional teams, including data engineers, transportation planners, and senior management, to ensure the effective use of data in transportation modeling.

Problem Overview

SANDAG's transportation models, particularly the EMME and OpenPaths simulation tools, were generating important insights about the region's traffic and transit flows. However, discrepancies existed between the outputs of two different versions of the EMME model (4.3.7 vs. OpenPaths 2024). Additionally, issues with the data quality—such as negative and zero values in key fields—were threatening the accuracy of the models. These challenges needed to be addressed in order to improve the decision-making process and enhance planning for future infrastructure development.

Approach & Solution

1. Data Comparison & Validation

The first step in improving the modeling process was to compare the outputs of Scenario 188 (EMME 4.3.7) and Scenario 206 (OpenPaths 2024) for the year 2022. Here's how I approached this task:

- **Data Extraction & Wrangling:** I extracted large datasets from the `abm_15_1_0_reporting` database, which contained model outputs for various transportation scenarios. The data included variables like ridership, route performance, and travel times.
- **Data Cleaning & Preprocessing:** Due to the volume of data and its complexity, significant data cleaning was required. This included filtering out irrelevant data points, handling missing values, and normalizing the data to ensure consistent formatting.
- **Model Comparison:** I compared the results of the two models, identifying areas where outputs significantly differed. I created detailed visualizations in PowerBI to show how the models differed, helping stakeholders identify which model was more accurate for certain transportation forecasts.
- **Findings & Insights:** I documented the differences between the two models, providing a comprehensive analysis of the performance of each scenario. This helped SANDAG decide which model to use for future transportation planning projects.

2. Data Quality Control & Automation

One of the critical challenges in the data I was working with was the presence of negative and zero values in the `BASEIVTT` column, which represented the in-vehicle travel time for different routes. These anomalies needed to be flagged and corrected to ensure that the modeling results were reliable. To solve this issue:

- **SQL-based Alerts:** I developed a set of automated alerts using SQL queries to detect when values in the `BASEIVTT` column were negative or zero. These alerts were designed to trigger warnings whenever such anomalies appeared in the data, preventing them from being processed further.

- **Data Integrity:** By implementing these alerts, I ensured that only valid data entered the analysis pipeline. This increased the overall quality of the models and improved the accuracy of forecasts related to transportation flows.

3. Visualization & Reporting

Once the data was cleaned and the models were compared, the next step was to communicate the findings effectively to key stakeholders:

- **PowerBI Dashboards:** I created interactive dashboards in PowerBI that visualized the differences between the two modeling scenarios. The dashboards highlighted key metrics such as trip ridership, route performance, and travel times, making it easier for decision-makers to grasp the implications of the findings.
- **Detailed Reports:** In addition to visualizations, I compiled comprehensive reports that explained the analysis and provided actionable recommendations. These reports were shared with transportation planners and senior management to guide future decisions on infrastructure and transit improvements.

4. Collaboration with Stakeholders

Throughout the project, collaboration was key to ensuring that the work was aligned with SANDAG's goals:

- **Cross-Functional Communication:** I regularly met with transportation planners, data engineers, and senior management to discuss findings, refine the analysis, and incorporate feedback. This collaboration ensured that the insights provided were directly applicable to ongoing transportation planning efforts.
- **Stakeholder Presentations:** I presented the findings to stakeholders in a clear, concise manner, using both visualizations and verbal explanations to make complex data more accessible. This helped to bridge the gap between data science and non-technical teams, ensuring that the results had real-world impact.

Tools & Technologies

The success of this project was enabled by a robust set of tools and technologies:

- **Databases:** PostgreSQL, MySQL, MongoDB for data extraction, management, and storage.
- **Languages:** Python (for data cleaning, analysis, and automation), SQL (for querying and data validation).
- **Visualization Tools:** PowerBI (for interactive dashboards and reporting), Tableau (for visualizing model comparison results).
- **Data Pipeline & Cloud Tools:** Databricks, AWS S3, AWS Glue (for data processing and pipeline management).

- **Automation Tools:** Jenkins (for automating data workflows), Apache Airflow (for orchestrating complex data pipelines).

Results and Impact

The outcome of this project had a significant impact on both the technical and operational sides of transportation planning at SANDAG:

- **Data Quality Improvements:** The automated alert system drastically reduced the occurrence of errors due to negative or zero values in the `BASEIVTT` column, enhancing the reliability of the transportation models.
- **Improved Decision-Making:** The comparison between the EMME and OpenPaths 2024 models provided actionable insights that allowed SANDAG to select the more accurate model for future transportation planning.
- **Stakeholder Engagement:** The interactive PowerBI dashboards and detailed reports enabled stakeholders to easily understand complex data, facilitating better decision-making processes for infrastructure development.
- **Process Optimization:** The automation of data validation and reporting processes streamlined the workflow, reducing manual intervention and speeding up the decision-making process.

Conclusion

My time at SANDAG as a Data Scientist was an invaluable learning experience that allowed me to apply advanced data science techniques to real-world problems in transportation planning. The work I did helped improve the accuracy of critical transportation models, while also enhancing data quality control and streamlining communication with stakeholders. I am proud of the work I accomplished at SANDAG and the impact it had on the region's transportation infrastructure planning.

Future Improvements

Looking ahead, I see several ways in which the work at SANDAG could be further enhanced:

- **Scalability:** The implementation of real-time data processing systems using AWS Lambda or Apache Kafka could allow for more immediate insights and updates to transportation models.
- **Predictive Modeling:** Incorporating machine learning models into the transportation planning process could improve predictions of future traffic patterns, helping SANDAG make more informed decisions about future infrastructure investments.
- **Cross-Agency Collaboration:** Expanding collaboration with other governmental and private agencies involved in transportation could lead to more accurate models and more effective policy recommendations.