COMP 496 PROJECT REPORT

This project is an extension of the previous data analysis that I performed on the dataset for the City of Calgary. While my earlier analysis provided invaluable insights into historical weather patterns through the exploration of 18,000 rows of data and the utilization of Python and PowerBI, this project takes a forward leap into the realm of weather prediction.

To create a robust weather predictions dashboard, I've seamlessly integrated multiple elements, each playing a crucial role in the predictive modeling process:

**1. Data Preprocessing:** I began by preparing the data, ensuring it's clean, structured, and ready for analysis.

**2. Model Evaluation and Selection:** Thoroughly evaluating and selecting the most suitable machine learning models became the cornerstone of this project, as accurate predictions depend on robust model choices.

**3. Feature Importance:** Understanding which weather variables have the most significant impact on predictions allowed for more informed modeling decisions.

**4. Hyperparameter Tuning:** Fine-tuning model hyperparameters was a vital step in optimizing predictive accuracy.

**5. Model Deployment:** I seamlessly transitioned from model development to deployment, making the predictions accessible to users in real-time.

**6. Creating a SQL Database:** The foundation of data storage was laid with the creation of a SQL database and the setup of a dedicated table to store data acquired through APIs.

**7. Data Fetching Function:** I designed a function that fetches real-time weather data from APIs and stores it efficiently in the SQL database.

**8. Column-to-Feature Mapping:** Ensuring alignment between the data stored in the SQL table and the variables used in the machine learning model was critical for accurate predictions.

**9. Query Design:** The creation of a well-defined query enabled seamless retrieval of data from the SQL database, facilitating the model's input.

**10. Dashboard Creation:** Leveraging the Dash library in Python, I crafted an interactive dashboard to visualize and present the prediction results effectively.

**11. Weather Code Categorization:** To enhance data interpretability, I implemented a function that assigns descriptive weather conditions based on weather codes.

**12. Predictions and Database Integration:** The predictive model was utilized to generate forecasts, with the results intelligently integrated back into the SQL database.

**13. Dash App Deployment:** I orchestrated the final deployment of the Dash app, constructing its HTML elements and ensuring its seamless functionality.

This project represents an exciting evolution of data analysis skills, enabling me to predict future weather patterns while seamlessly incorporating multiple elements of data management, machine learning, and visualization.

In the coming sections, there is a detailed report on each of these elements and how they have come together to form a powerful weather prediction dashboard.

**1. Data Preprocessing**

**x = calgary\_weather\_df[['relativehumidity\_2m (%)', 'apparent\_temperature (°C)', 'surface\_pressure (hPa)', 'precipitation (mm)',**

**'snowfall (cm)', 'cloudcover (%)', 'windspeed\_10m (km/h)']]**

**y = calgary\_weather\_dfy = calgary\_weather\_df[['temperature\_2m (°C)', 'weathercode (wmo code)']]**

**from sklearn.model\_selection import train\_test\_split**

**x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.2, random\_state=42)**

In the realm of data science and machine learning, effective data preprocessing is a fundamental step in ensuring that data is well-structured and suitable for model training.

In this project, data preprocessing played a pivotal role in preparing the dataset for predictive modeling. Here are the technical details of the data preprocessing steps undertaken:

Feature Selection:

The initial dataset, represented by the variable **calgary\_weather\_df**, contained various weather-related variables. From this dataset, a subset of relevant features was selected for modeling. These features included 'relativehumidity\_2m (%)', 'apparent\_temperature (°C)', 'surface\_pressure (hPa)', 'precipitation (mm)', 'snowfall (cm)', 'cloudcover (%)', and 'windspeed\_10m (km/h)'.

Target Variables:

Two target variables of interest were identified for prediction: 'temperature\_2m (°C)' and 'weathercode (wmo code)'. 'temperature\_2m (°C)' represents the target variable for temperature prediction, while 'weathercode (wmo code)' was chosen for weather code prediction.

Data Splitting:

* To facilitate model training and evaluation, the dataset was divided into training and testing subsets. The **train\_test\_split** function from the **sklearn.model\_selection** module was employed for this purpose.
* A common practice in machine learning, an 80-20 split ratio was chosen, with 80% of the data allocated for training (**x\_train** and **y\_train**) and 20% for testing (**x\_test** and **y\_test**).
* The inclusion of a **random\_state** parameter set to 42 ensures reproducibility, as it seeds the random number generator for consistent data splits.

This data preprocessing step lays a solid foundation for subsequent model development and evaluation. It involves the selection of relevant features, identification of target variables, and partitioning of the dataset into training and testing sets. These steps collectively prepare the data for training predictive models to forecast temperature and weather codes accurately.

2. **Model Comparison and Selection:**

In this phase of the project, I conducted a comprehensive evaluation of three distinct machine learning models for predicting weather variables based on the preprocessed dataset. The models that I examined were Linear Regression, XGBoost, and Random Forest. The assessment encompassed several performance metrics to determine which model best aligns with the predictive objectives.

**Linear Regression:**

* RMSE (Root Mean Squared Error): 6.8777
* MSE (Mean Squared Error): 87.3486
* MAE (Mean Absolute Error): 4.1339
* R-squared: 0.7931

**XGBoost:**

* RMSE: 0.1739
* MSE: 0.0402
* MAE: 0.1001
* R-squared: 0.9997

**Random Forest:**

* RMSE: 0.2687
* MSE: 0.0790
* MAE: 0.1348
* R-squared: 0.9995

**Model Analysis Summary:**   
The performance evaluation revealed insightful results. Linear Regression, while a valuable baseline, displayed notably higher error metrics compared to XGBoost and Random Forest. Thus, Linear Regression was ruled out as the primary predictive model.

Between XGBoost and Random Forest, a closer examination of the metrics indicates a significant difference. XGBoost showcased superior performance across all four metrics—RMSE, MSE, MAE, and R-squared. Its lower RMSE, MSE, and MAE, coupled with the near-perfect R-squared value, collectively suggest that XGBoost excels in capturing the intricate relationships within the weather dataset.

Therefore, the selection process concluded that XGBoost stood out as the optimal predictive model. Its exceptional performance metrics and robust predictive capabilities position it as the model of choice for future weather predictions. This model will be utilized in subsequent stages of the project for its remarkable predictive accuracy.

3.

Top of Form

Bottom of Form