Mini-project 2024

INFT 2003 Big Data

Task 1

a)

- In Big Data we can find the three Vs: *Volume*, *Velocity* and *Variety*. Volume in big data is defined as large amounts of data, from size of terabytes and zettabytes in the dataset. Velocity is the speed of the data, in how they are generated, analyzed and gathered. This includes batches, real-time and streams. The last V, Variety, is the different datatypes. Data can come from an internal or external data source, or as structured, semi-structured or unstructured data (Xiaomeng Su, "*Introduction to Big Data*", p.2).

  The article examines various data from the Low Carbon London (LCL) project, which collected household energy consumption using smart meters installed in London households.

  Smart Meter Data:

  - *Volume*: over 167 million records were collected from over 5000 households in 30-minutes intervals from November 2011 to February 2014.
  - *Velocity*: Data is collected at 30-minutes intervals, making it a high velocity data stream requiring near real-time processing.
  - *Variety*: highly structured data, detailing every energy usage value for each household.

  Bank Holiday Data:

  - *Volume*: a smaller dataset containing information about bank holidays in the data collection period.
  - *Velocity*: the data is static and doesn't update often, therefore low velocity.
  - *Variety*: structured data providing information about a day and if it is a bank holiday or not.

  Weather Data:

  - *Volume*: multiple weather parameters over the same period collects lots of data.

- *Velocity*: the weather data is retrieved from APIs, which is often updated in real-time for quick decision making.
- *Variety*: semi-structured data, containing data about humidity, temperature etc.

Demographic Data:

- *Volume*: a smaller volume but contains data of each household's socioeconomic status (ACORN categories).
- *Velocity*: the data is static and not updated frequently.
- *Variety*: structured, categorical data classifying households in socioeconomic groups.

- Load Forecasting in smart grid is a Big Data problem because it involves handling large volumes of diverse data, real-time streaming, and the need for scalable analytics tools. The three Vs for Load Forecasting in smart grid:

  Volume: The large amount of data generated by the smart meters in the study is one of the biggest reasons why Load Forecasting becomes a Big Data problem. In the LCL project, in total there were over 167 million records of the total readings, collected by 5,567 households. Traditional data processing tools would not be able to handle this volume of data efficiently.

  Velocity: The data from the smart meters need to be accessed in real-time or near-real-time processing to support dynamic operations. The data is generated continuously at a high speed, and quick decisions needs to be made, therefore it is a Big Data problem.

  Variety: The data generated is not only limited to consumption readings, but it also contains weather data, social and demographic information, as well as geographic data. The mix of structured data (e.g. energy consumption data), semi-structured data (e.g. weather data) and unstructured data (e.g. SMS notifications) increases the complexity of processing and analysis.

b)

- The business problem the Load Forecasting application this case study is addressing is accurately predicting energy consumption. This includes balancing the energy demand and supply and to optimize smart grid operations to make it more economical and efficient.

- A potential business benefit of a good Load Forecasting is to decrease cost. By predicting energy consumption, demand will be met, and it will be easier to allocate resources. It will also help avoid grid overload, and secure operational efficiency. Another benefit is customer satisfaction by enabling tailored tariffs and electricity supply.

- **Supervised and unsupervised learning**:
  In supervised learning we have a dataset that includes the target variable. It is used to learn a model to predict a target variable based on historical data, which can be used to make new predictions on new cases. Unsupervised learning on the other hand is more explorative and produces groupings based on assumptions without a target variable to test its findings on (Xiaomeng Su, "*Predictive Analytics*", p. 15). Since the LCL data has a target variable (e.g. "the sum of the energy consumption of all residential houses"), a supervised learning method is the best approach.

  Tasks for Load Forecasting:
  Commonly used models in supervised learning are regression models and classification. The two main uses of regression models are:

  - *Prediction*: estimate an outcome on historical data.
  - *Inference*: find relationships between features and their influence on the target variable.

  To evaluate the regression models, R^2 and the Mean Squared Error (MSE) is often used.

Classification is frequently used to group or classify data based on patterns learned from historical data. To estimate the accuracy of the classification model, a confusion matrix is made. It is used to measure the proportions of actual correctly predicted values and gives an indication of the level of confidence you should have in the model (Xiaomeng Su, "*Predictive Analytics*", p. 10-11).

The target variable in the Load Forecasting application is "the sum of the energy consumption of all residential houses", which is a continuous numerical value. Therefore, a regression model is the better choice.

c)

- The author proposed different descriptive analytic figures: histogram, scatter plot, time-series, bar chart etc. Another descriptive analytic figure that could be utilized in this article is a time-series plot of energy consumption during holidays.

  - *x-axis*: days leading up to, during, and following specific holidays (e.g. Christmas, easter, new year's etc.)
  - *y-axis*: energy consumption (kWh)
  - *Groups (optional)*: ACORN categories (e.g. Affluent, Comfortable, or Adversit)

  This visualization would illustrate how energy consumption patterns vary around different holidays. It could help identify specific holiday-related behaviors, such as increased energy consumption during festive holidays. By adding the ACORN categories, you could get a better insight in how the different demographic and socioeconomic groups behave during the holidays as well.

- **Multivariate Polynomial Regression (MPR):**
  MPR is an extension of Linear Regression and is used to model a target variable with multiple independent variables, capturing more complex relationships between features (Saturn Cloud, 2023).

Capturing non-linear relationships:

As already stated, an advantage of using MPR is getting a better understanding of non-linear relationships in the dataset, compared to a simple linear regression only capturing linear relationships (Saturn Cloud, 2023). This is important for the smart grid domain because energy consumption datasets often are influenced by a wide range of non-linear data (e.g. weather conditions, time of day and customer behavior). For instance, energy consumption may peak at holidays, or during specific seasons, these trends are rarely linear. MPR allows us to get a better understanding of these relationships and predict complex consumption patterns with great accuracy.

Flexible in adjusting model complexity:

MPR also uses polynomial equations to adjust the complexity of the model, making it more flexible in fitting intricate patterns in the data (Saturn Cloud, 2023). Energy consumption datasets often have various and intricate patterns, often influenced by multiple interacting factors. A flexible model like MPR is essential for effectively modeling these highly complex relationships.

Easy to implement and familiarity:

Several libraries in python have existing models built in (e.g. scikit) making the model easy to use and implement. Since MPR is an extension of the linear regression framework it makes it easier to use and implement for users already familiar with linear regression (Saturn Cloud, 2023). As discussed earlier, smart grid is a Big Data problem and has large volumes of data that needs to be processed with high velocity, preferably in real-time. In this context, having efficient and easy-to-implement models is critical. Many professionals are already familiar with linear regression, making it easier to implement and handle the MPR.

Decision Tree Regression (DTR):

DTR is used for predictive modeling and can be used to assist in decision making. It works recursively by dividing the datasets into smaller subsets based on the target features values (Alexander Holt, 2020).

Easy to interpret:

DTR is easy to interpret because they split their data based on decision rules that are easy to visualize and understand, also for non-technical stakeholders (GeeksForGeeks, 2024). This is important in smart grid to allow stakeholders to easily understand and read the predictions being made and helps to trust the predictions being made.

Capturing non-linear relationships:
Unlike simple linear regression, DTR can represent complex relationships between features and the target variable (GeeksForGeeks, 2024). Energy consumption patterns are often dependent on weather, time and demographics, and DTR will help find non-linear relationships between these features.

Can handle different data:
The DTR models are not sensitive to scaling, and therefore there is no need for feature scaling. The model can also handle variations of data types, such as numbers and categories, making it easy to implement to datasets with mixed datatypes (GeeksForGeeks, 2024). The smart grid dataset contains variations of datatypes (e.g. smart meter data and SMS notifications) making the DTR a good fit for an easy implementation without using hot encoding etc. It also eliminates the use for feature scaling on all the mixed datatypes.

- The authors performed feature selection based on the results from the Descriptive Analytics. Feature selection helps eliminate irrelevant or redundant features, which can otherwise cause noise in the dataset. This not only simplifies the model, but also improves generalizability by reducing the risk of overfitting. The use of regulated linear regression models further reinforces the approach, reinsuring that the models focus on the most important features while maintain a balance between bias and variance (Ali El-Sayed, 2024, p.8).

- **Explanation of evaluation metrics:**
  R^2: a high R^2 score (closest to 1) indicates the proportion of variation that are explained by the model. If the score is equal to 1, the model has a perfect prediction (Xiaomeng Su, "*Predictive Analytics*", p.13).

Mean Absolute Error (MAE): the MAE is calculated as the absolute difference between the actual and the predicted values in the dataset. A low MAE indicates a smaller difference between the predicted and the actual values, making the model a good fit. (Medium, 2020).

Mean Squared Error (MSE): MSE is the average of the squared difference between the predicted and the actual values in the dataset. A low MSE indicates a small difference and a good fit of the model (Medium, 2020).

Root Mean Squared Error (RMSE): a low RMSE score indicates a good fit of the model. It is calculated by taking the square root of the MSE score (Xiaomeng Su, "*Predictive Analytics*", p.13).

"Only one block"

Looking at the results in Table 6 (Ali El-Sayed, 2024, p. 19) the best model for "only one block" was plain regression. The plain regression model for the "only one block" had the highest R^2 score of (0.944), lowest MAE (1.721), lowest MSE (6.112) and lowest RMSE (2.475).

"All houses"

The best model for "all houses" is close between the Ridge Regression and the Elastic net with the same highest R^2 (0.965), lowest MAE (68.073) and lowest RMSE (96.117) scores. There is only a difference in the MSE score (Ridge Regression: 9238.572 and Elastic net: 9238.573), where the Ridge Regression has a slightly lower score. There is a very small difference, but the Ridge Regression is the best model based on the slightly lower MSE score.

d)

- Data Analytics proposal on Load Forecasting at TrøndEnergi using the CRISP (Cross Industry Standard Process for Data Mining) framework. The project will be divided in six major phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment (Xiaomeng Su, "*Process, Team and Privacy*", p. 2).

Business understanding

The business problem TrøndEnergi aims to resolve is to *ensure efficient grid operations while integrating renewable energy sources*.

This problem requires accurate Load Forecasting to correctly balance demand and supply and optimizing energy distribution. By achieving this, TrøndEnergi can accomplish:

- *Cutting operational costs*: minimize energy waste.
- *Increased grid reliability*: prevent grid failures during peak loads.
- *Sustainability*: optimize renewable energy usage.

The primary task involves using regression models and supervised learning on historical data, preferably from the last 12 months. This is used to predict continuous load values.

Data understanding

The data used in this business problem includes historical energy consumption from smart meters, real-time weather data and customer demographic information.

Smart meter data: 30 minutes intervals energy usage data for households and businesses, aggregated by region to capture broader trends.

Weather data: temperature, humidity, wind, etc. collected historically and in real-time.

Demographic data: household size, city, economy status etc.

Some challenges may arise in data collection:

- Missing or inconsistent values from the smart meter data in case of a malfunction.
- Combining structured data like smart meter readings with semi-structured data (e.g. weather conditions) needs considerate and careful handling.

- The big volume of historical and real-time data needs a robust infrastructure for storage and processing.

To address these challenges, the data will aim for completeness, accuracy, and timeliness, ensuring it aligns with the business problem.

Data preparation

As part of the ETL (Extract, Transform, Load) process, the data should be cleaned of missing values and remove irrelevant data points. The features may also be scaled, if relevant, to ensure a uniform representation of all features.

The dataset will be split in training and testing sets to ensure robust evaluation of the models.

To ensure a secure level of privacy, all data will be anonymized using tokenization techniques.

Modeling

In the modeling phase we will evaluate multiple regression models, including Decision Trees, Multivariate Polynomial Regression and Ridge Regression, to determine the best approach for accurate Load Forecasting.

Evaluation

The models will be evaluated on metrics, such as R^2, MSE and MAE. The evaluation will help provide a fitting model for this business problem. A successful model will demonstrate strong performance on historical data and generalize well to unseen data. We will have review sessions with stakeholders to ensure the models align with their requirements.

Deployment

The deployment phase will implement the model as part of the TrøndEnergi grid systems.

Real-time updates: the model will be used to predict load dynamically and add grid operations accordingly.

Dashboard improvement: create a user-friendly interface to visually display the Load Forecasts.

Periodic updates: periodically retrain the model on new data to ensure consistent accuracy in load predictions and adapt to new trends.

By implementing this proposal, TrøndEnergi can achieve substantial operational improvements and strengthen its position as a leader in sustainable energy management.

References:

1. Xiaomeng Su. *Introduction to Big Data*. Lecture. NTNU.

2. Xiaomeng Su. *Predictive Analytics*. Lecture. NTNU.

3. Xiaomeng Su. *Process, Team and Privacy*. Lecture. NTNU.

4. GeeksForGeeks (2024). *Pros and Cons of Decision Tree Regression in Machine Learning*. Web article. https://www.geeksforgeeks.org/pros-and-cons-of-decision-tree-regression-in-machine-learning/

5. Meduim (2020). *MAE, MSE, RMSE, Coeffecient of Determination, Adjusted R Squared – Which Metric is Better?*. Web article. https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

6. Alexander Holt (2020). *Decision Trees*. GitLab. https://gitlab.stud.idi.ntnu.no/alexholt/python-big-data/-/wikis/Decision%20Trees

7. Saturn Cloud (2023). *Multivariate Polynomial Regression with Python*. Web article. https://saturncloud.io/blog/multivariate-polynomial-regression-with-python/#pros-and-cons-of-multivariate-polynomial-regression

8. Ali El-Sayed (2024). *Big data resolving using Apache Spark for load forecasting and demand response in smart grid: a case study of Low Carbon London Project*. Journal of Big Data.