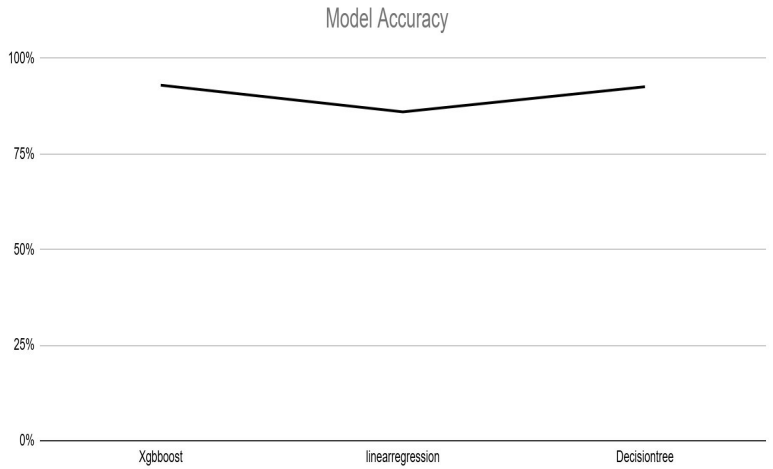
Iron Kaggle

G1

Mehak
Jurgen
Sven

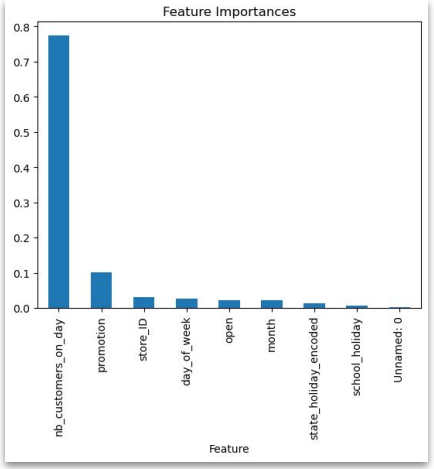
Executive summary

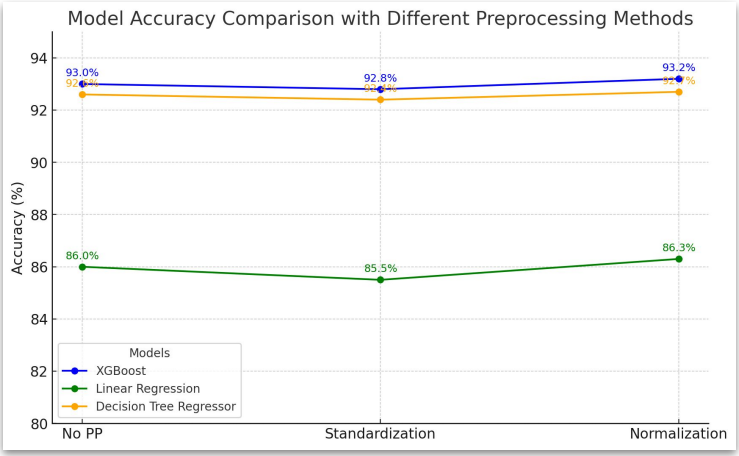
- Accuracy with training data (sales.csv): 93%
- Best model: xgboost regressor
- R^2 Prediction: 93%
- Quick recap of alternatives considered:
  - LinearRegression 86%
  - DecisionTreeRegressor 92.6%



Methods (preprocessing)

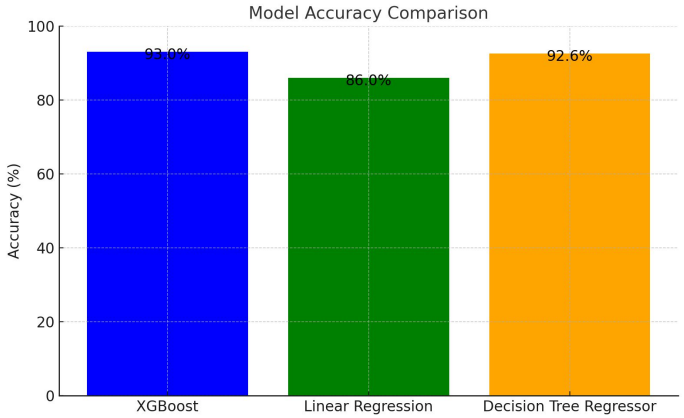
- Mapped and converted values for the columns ‘date’ and ‘state_holiday’
  - ‘Date’ -> convert to datetime object and extract the month
  - ‘State_holiday’ -> used label encoding
- Plotted feature importance
- Removed the unnamed column of ID values
- Tried out log transformation but without any benefit
- Split the data into train and test groups





Methods (models) – 1 or 2 slides

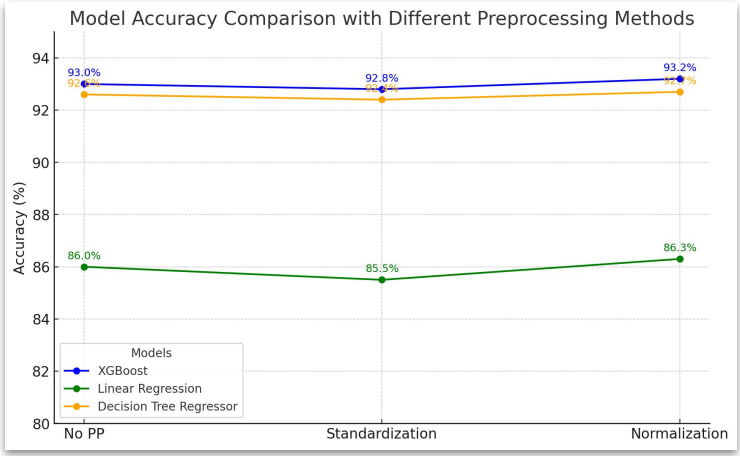
- Since we need to predict numerical sales values we have selected:
  - XGBoost Regressor
    - Prevents overfitting and handles large datasets
    - Commonly used in forecasting sales
  - Linear Regression
  - Decision Tree Regressor



Selected model: XGBoost Regressor

- Initial r2 score: 93%
  - After standardizing feature data: 93%
  - After normalizing feature data: 93%

- Neither normalization nor standardization improved the accuracy of our model



Takeaways

- Recap / conclusions
- Challenges
  - How to handle the high correlation between nb_customers_on_day and sales?
  - Balancing model complexity with model performance
- Key learnings
  - Calculate the accuracy after making bigger changes to the data (e.g. dropping columns) to be able to check if the changes improve the accuracy

Instructions

- You can make the charts with python or excel.
- All team members must **participate** (either split the slides, or discuss the part that you did for each slide)
- **7 minutes maximum** for presentation + 3 minutes for questions
  - 7 is a HARD limit. Aim for 5, it should be enough.
- Tip: **Rehearse** the presentation at least once