Architectural Decisions Document

By: Alexander Stetzer

1 Data Source

1.1 Technology Choice

The data for this project comes from a kaggle dataset on water pumps. The data will come as a csv file for easy loading.

1.2 Justification

The kaggle dataset was chosen as it was a free data source and the water pump data had many timestamps.

2 Enterprise Data

2.1 Technology Choice

No Enterprise Data is needed for this project.

2.2 Justification

The reason that no Enterprise Data is needed for this project is that subsets of the data do not need to be uploaded to the cloud

3 Streaming analytics

3.1 Technology Choice

The current model will not have streaming analytics involved but the deep learning model could be transformed into a detector

3.2 Justification

Streaming Analytics were not used for this as the goal was to see if a model could be made to predict the current status of the water pump

4 Data Integration

4.1 Technology Choice

The data will be read into a jupyter notebook as a csv file. From there it will be transformed into a pandas DataFrame where the features will be made. Any missing values will be filled with sensor means. Standard Scaling will be used to help wrangle the data.

4.2 Justification

Pandas is used for the project as it is a go to for data science. The mean of each sensor will be used to replace any missing value as this will have the smallest impact on the final project and it is better than completely removing or placing zeros as the missing data. Standard Scaling will be used as the data has large differences in ranges.

5 Data Repository

5.1 Technology Choice

The Data Repository for the project is the IBM Object Store.

5.2 Justification

The IBM Object Store is used for this project because it is a free storage system for the data. Also the Object Store will allow for ease of loading the data into the notebook and saving the data from the notebook.

6 Discovery and Exploration

6.1 Technology Choice

To discover and explore the data, I will check means and make boxplots to visualize the important data in the set. Checking the null values will also be done and checking the unique values for the machine status.

6.2 Justification

Using the means and boxplots to check the data is important. Being able to see outliers or potential odd data is important to make sure the data is good. Checking for null values is also important as null values could confuse the model and the null values could potentially hurt the model performance. The unique values of the machine status are important to find because in order to change the data type of the prediction column the actual values need to be known.

7 Actionable Insights

7.1 Technology Choice

For the deep learning model I am going to use Tensorflow with Keras. RandomizedSearchCV will be used to check the best hyperparameters. For the non deep learning model I am going to use the sci-kit learn support vector machine, SVC. Instead of RandomizedSearchCV, GridSearchCV will be used here. Both of the models will be assessed with F1-score and Matthew's Correlation Coefficient.

7.2 Justification

Using tensorflow is a no brainer here as it is a very well understood and great deep learning product. RandomizedSearchCV is used to find the best hyper parameters as the hyperparameters here are continuous. SVC is used for the non deep learning model as it is fast and a very accurate model for binary classification. GridSearchCV is used here because the hyperparameters used are discrete and there are only a few different options to choose from, resulting in low training time. F1-score is used to check the models as it is a well known accuracy measurement and can be easily interpreted. Matthew's Correlation Coefficient is used as well because it takes true negatives into its accuracy calculation. Taking true negatives into account is important as the goal is to make sure that all true negatives are found.

8 Applications / Data Products

8.1 Technology Choice

The application for this project is that it can be converted into a detector for the pump. The Data Product for this project is a PDF of both model notebooks.

8.2 Justification

Since the use case is to check if the water pump sensor data can be used to predict the actual state of the water pump, a notebook with markdown explanations is only needed.

9 Security, Information Governance and Systems Management

9.1 Technology Choice

I will be the only one who can access the notebooks and data. Also all credentials will be in hidden notebook cells with other markdowns to explain what the hidden cells do.

9.2 Justification

Since I am the only one that needs to access the information the hidden cell method will be sufficient to protect the information.