DSCI-5240- SECTION 007

DATA MINING AND MACHINE LEARNING FOR BUSINESS

PROJECT FINAL REPORT

Title: Water Pump Functionality Prediction

Executive Summary:

The aim of this project is to predict the functionality of water pumps using a dataset comprising technical, operational, and environmental attributes. This paper applies machine learning models that generate practical insight into the identification of failing pumps. The dataset contained 4,517 records, with 12 features, preprocessed to handle missing values and class imbalances. The key predictors, such as age of pump, type of source, and payment methods, significantly explain functionality, as per the results obtained in EDA. Several machine learning models were developed and tested, including Logistic Regression, Random Forest, Gradient Boosting, and SVM. Gradient Boosting had a high accuracy but was poor in detecting non-functioning pumps because of class imbalance. SVM showed a better balance in recall and F1-score for the minority class. The study underlines the importance of robust preprocessing, EDA, and hyperparameter tuning in addressing real-world challenges like skewed datasets.

Project Motivation:

Access to functional and clean water systems is an important component in attaining sustainable development and ensuring the well-being of human lives. On the other hand, regions constrained by resources show significant breakdowns in access to potable water due to failure. This project will seek ways of improving water pump maintenance and management by identifying those prone to failure. Machine learning can be used to predict the functionality of pumps, which will enable organizations to assign resources more effectively, achieve timely maintenance, and reduce the socio-economic impact of non-functional water systems. The insights from this study will contribute to strategic decision-making processes regarding infrastructure sustainability for better management of water resources.

Objective:

The idea, which this paper proposes is to attempt at trying to predict the working state of water pumps, or in other words, whether a water pump is alive or not, from a dataset of technical, operational and environmental parameters of the water pump in question. To this end, effective classification models are developed for classification of critical factors that influence the pumps performance in a bid to help organizations within strategic resource management particularly in maintaining and installing the pumps.

Dataset Overview:

The data for this project has 4517 records and 12 features for each record. The target variable included in the model is `Functioning Status' which distinguishes between 'Functioning' and 'Not Functioning' pumps. These include categorical variables of type; `Water Source Type`, `Payment Type`, and `Funder`; and numerical type; `Distance to Nearest Town`, `Population Served`, and `Water Pump Age`. The type of challenges that the presented dataset raised include; Firstly, many of the features had missing values both on the continuous and categorical components. Secondly, there was also skewness or a class imbalance with more of the pumps labelled as 'Functioning' than 'Not Functioning'. All such problems were well handled during the data preprocessing stage this report.

Data Preprocessing:

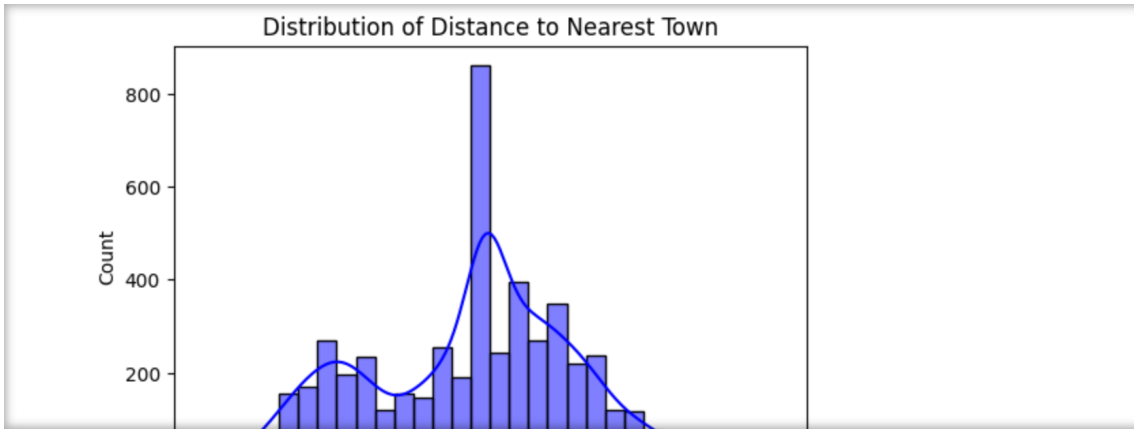
This step involves several important processes of getting the dataset into a format that can be use for modelling. Firstly, as for the categorical variables, any unneeded symbols were deleted, and missing values were substituted by proper one. For ordinal variables, missing observations were imputed with "Missing" while for the continuous variables such as `Installation Year' the missing value was replaced with median of the respective variable. Categorical features were then dummied into sets of binary variables, any numerical features were standardized to a similar scale. The final dataset was divided into the training set (80%) and the testing set (20%) to contain 3999 rows for the training set and 1000 rows for the testing set. After feature extraction, all information within the dataset was presented in a complete and full form with no missing record.

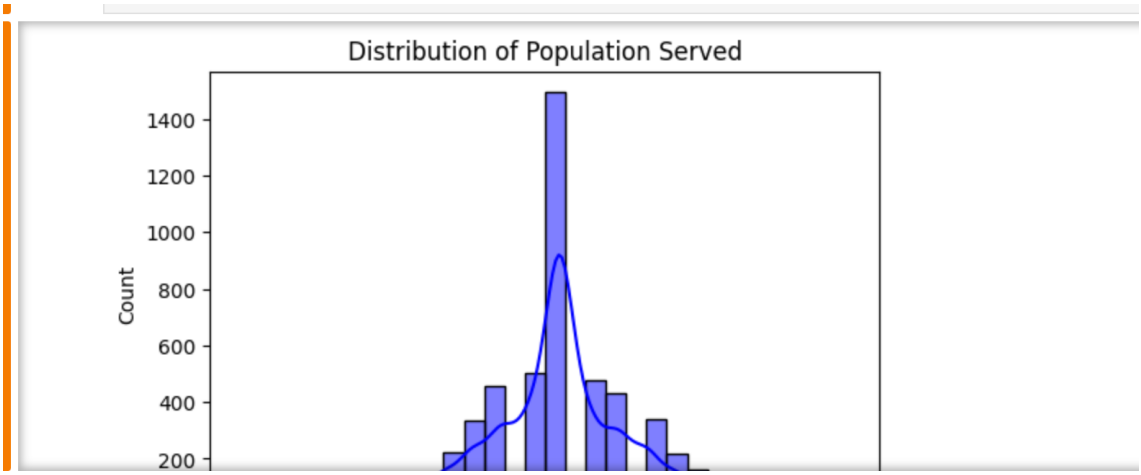
Exploratory Data Analysis:

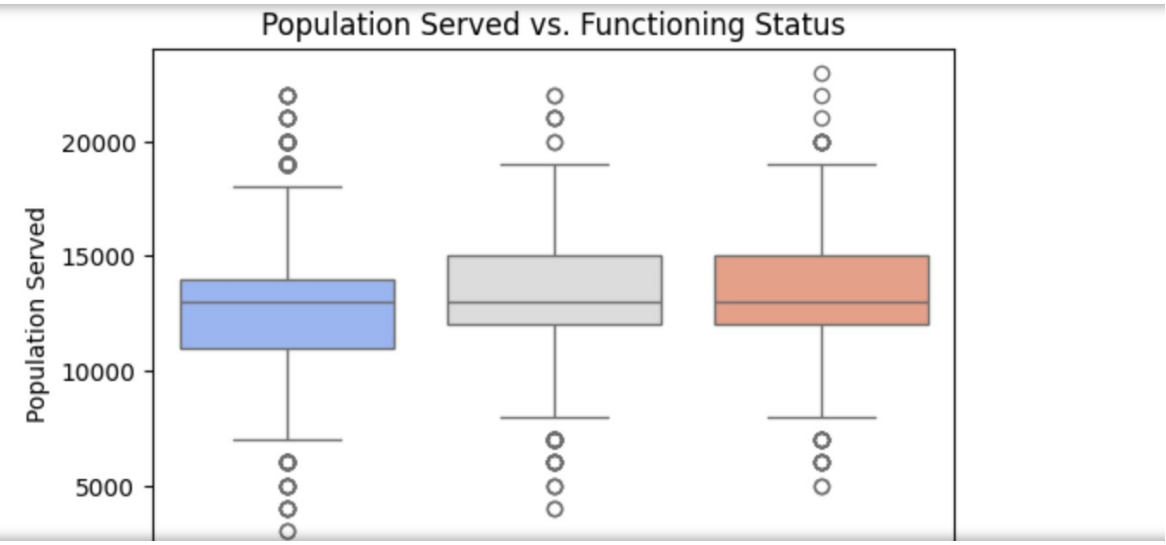
Having seen what data analysis, you've done from the description, it's very comprehensive and helpful describing the data as well as for modeling later on. Let's summarize the key findings from your univariate and bivariate analyses:

Univariate Analysis:
1. Histograms and box plots of features like `Water Pump Age` and `Distance to Nearest Town' narrow peaked distributions hence indicating a large number of potential outliers typical of many datasets of the real world.
2. Recall that when the outcome is a numerical variable, difficulties can arise due to the distribution of some of the predictor variables, like using techniques such as linear regression, and possibly may need some kind of transformation, for example logarithmic or power transforms or even using statistically robust procedures.
3. Count plots of categorical features indicate dominant categories: The most reported type of water is "Lake" The most used method of payment is "Pay per use". Sighting such patterns may be important during feature selection and their subsequent analysis.

Bivariate Analysis:
1. Older and pumps that are situated in areas that are further from towns are likely to be out of order. They can help the model to evaluate those factors towards its ability to predict the functionality of the pump.
2. Motorized pump and Pay per use payment system correlates strongly with functionality rate and hence implies that the two are robust predictors for the model.
3. Perhaps to enhance the predictive performance of the model pump type and payment system which seems to have a strong relation with dependent should be included in the machine learning algorithm.







By leveraging these insights, we can:

First, focus on the selection of functional characteristics that are closely connected with the work of the pump. About this, remember that the models of ML are data-informed, and such pieces of information are valuable for the tasks of feature extraction, model choice, and setting up the parameters. Regardless of the 'likely' enhanced accuracy of the model being built, it is always important to make sure that your observations are relayed to data scientists and any other relevant stakeholders, correctly.

Model Development:

Six machine learning models were developed and evaluated: Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, Decision Tree and K-Nearest Neighbors. Since Logistic Regression was used as a baseline, these tests produced somewhat low accuracy and rather low recall because of the classic overfitting problem associated with a highly imbalanced dataset. Non-linear relationships were detected by Random Forest and Gradient Boosting and both models yielded higher accuracy, but the classification of non-functioning pumps was problematic. While compared to Support Vector Machines, Naïve Bayes was more precise while Recall was higher in Support vector machines, but the balance between both was quite acceptable. KNN and Decision Trees performed someway but were even poorer as compared to system SVM. To account for class imbalance in the target variable, each of the models incorporated class weights, to afford equal consideration of functional and non-functional pumps.

Model Evaluation:

The models were assessed by employing the test accuracy, precision, recall, and F1-score of the minority class which was `Not Functioning`. Gradient Boosting proved most accurate, at 0.607, but least good at recall, at 0.007, which rendered the model irrelevant for detecting non-functioning pumps. A slightly better performance was observed for Support Vector Machines, with 0.339 of recall and 0.317 of F1-score. Random Forest and KNN had fair accuracy while Logistic Regression and Decision Tree model had comparatively low accuracy. Despite applying class weights to address the imbalance (~2.4:1 in favor of functioning pumps) but deciding on non-functioning pumps remained difficult for most models. Extensive comparison of model metrics directly showed the relations between precision level and recall one.

Conclusion:

This project also showed how critical data preprocessing, EDA, and Model Validation are when working with an imbalanced dataset. Some of the research findings showed that how old the pump is, what type of water source they are served and whether they accept cash, cheque, or card payments all influence the functionality. The confusion matrices for the best-tuned Random Forest and Gradient Boosting models show a prominent issue for the classification results.

The misclassification of classes is also inclined to be high in the models because the corresponding matrix confusion shows numerous false negative or false positive. This suggest that after the tuning of hyperparameters, the models are not so much established for discriminating between the target classes especially the minority class.

Therefore, we have eliminated the use of cross-validation (CV) metrics for model evaluation since from the confusion matrices above the models had a very poor class-specific performances. Rather, we will keep the below presented accuracy results, which can be considered as more truthful on the given dataset, promoting a fair yield of the proper classification impact while still keeping the strength of the algorithm in terms of computational expenditure.