PUMP IT UP!
DATA MINING THE WATER TABLE

Data Science for Social Good

HOW CAN WE APPROACH THE WATER SCARCITY IN TANZANIA ?

METHODOLOGY

MACHINE LEARNING TECHNIQUES CAN HELP OUT

Data

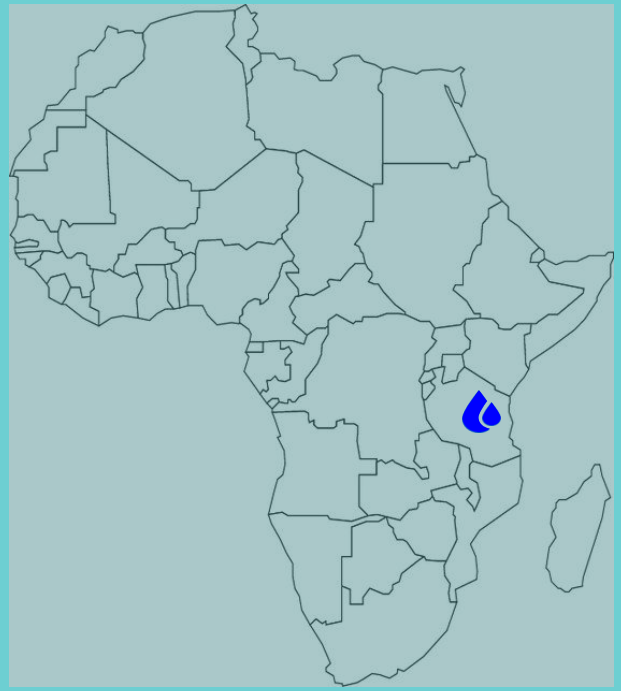
Preprocess

Machine Learning

Decisions

PROBLEM DESCRIPTION

TANZANIA WATER AND SANITATION CRISIS



ABOUT 59 MILLION

40 % URBAN
60 % RURAL

50 MIN ROUND TRIP TO ACCESS WATER POINTS

25 MILLION LACK OF SAFE ACCESS TO SAFE WATER

NO PROPER SANITATION → WATERBONE DISEASES

DATA PREPROCESSING

WHAT DATA DO I HAVE AND WHAT ARE THE NEXT STEPS ?

STARTING POINT

59400 WATERPOINTS
41 FEATURES:
GEOGRAPHICAL
TECHNICAL
DEMOGRAPHIC

STEPS

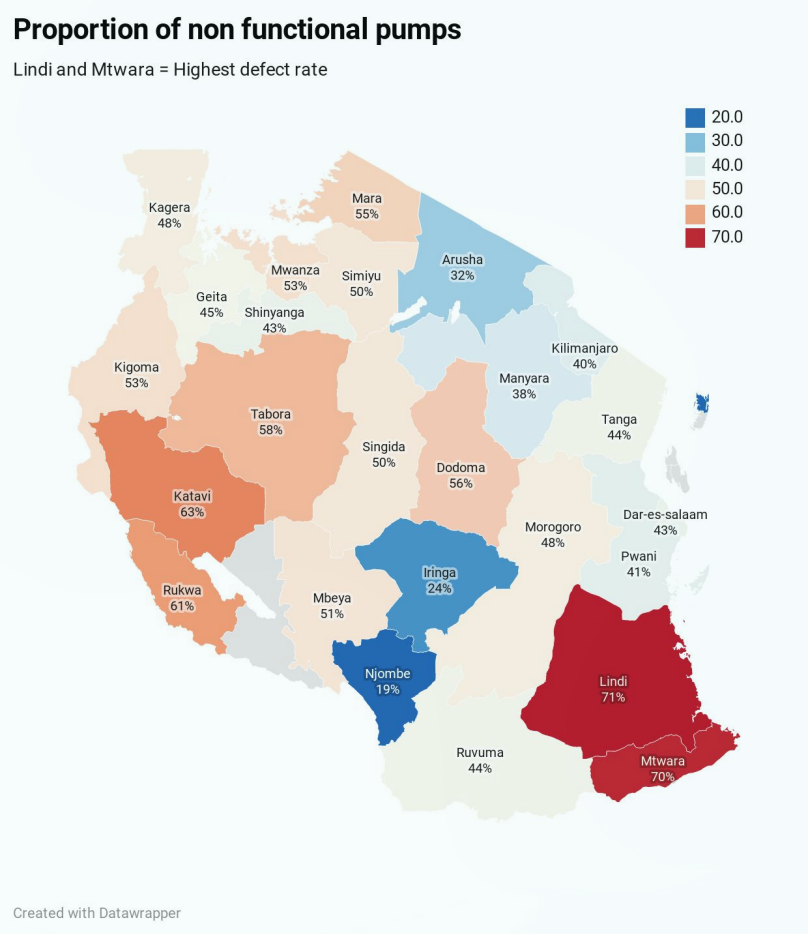
DATA CLEANING
DATA EXPLORATION
FEATURE ENGINEERING AND MACHINE LEARNING

TARGET

WATER PUMP FUNCTIONALITY
FUNCTIONAL
NON FUNCTIONAL
FUNCTIONAL - NEEDS REPAIRS

DATA INSIGHTS

MOST REGIONS HAVE >50 % NON FUNCTIONAL WELLS!!

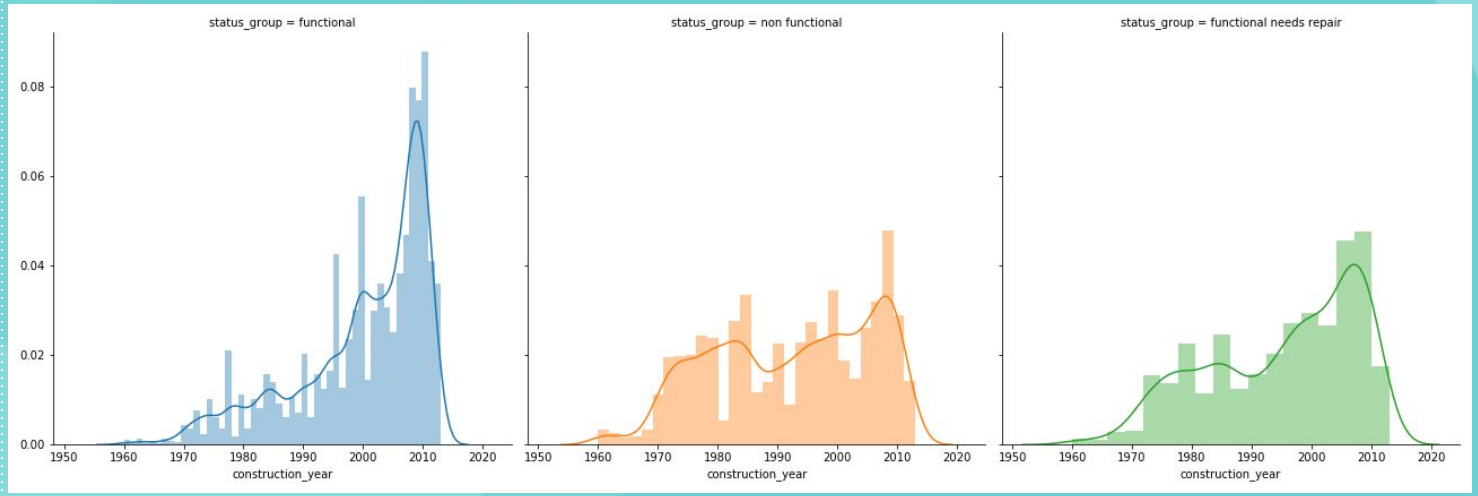


LINDI AND MTWARA LESS FUNCTIONAL REGIONS

NJOMBE IRINGA MOST FUNCTIONAL REGIONS

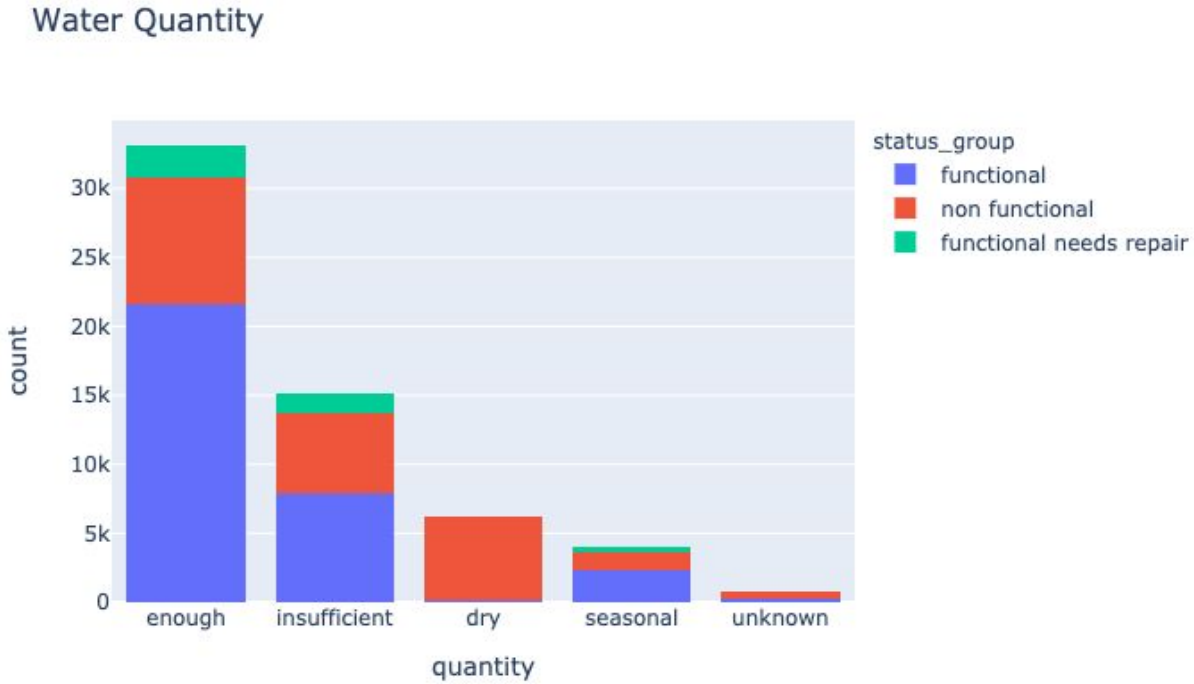
DATA INSIGHTS

WATER PUMP AGE DEFINES FUNCTIONALITY



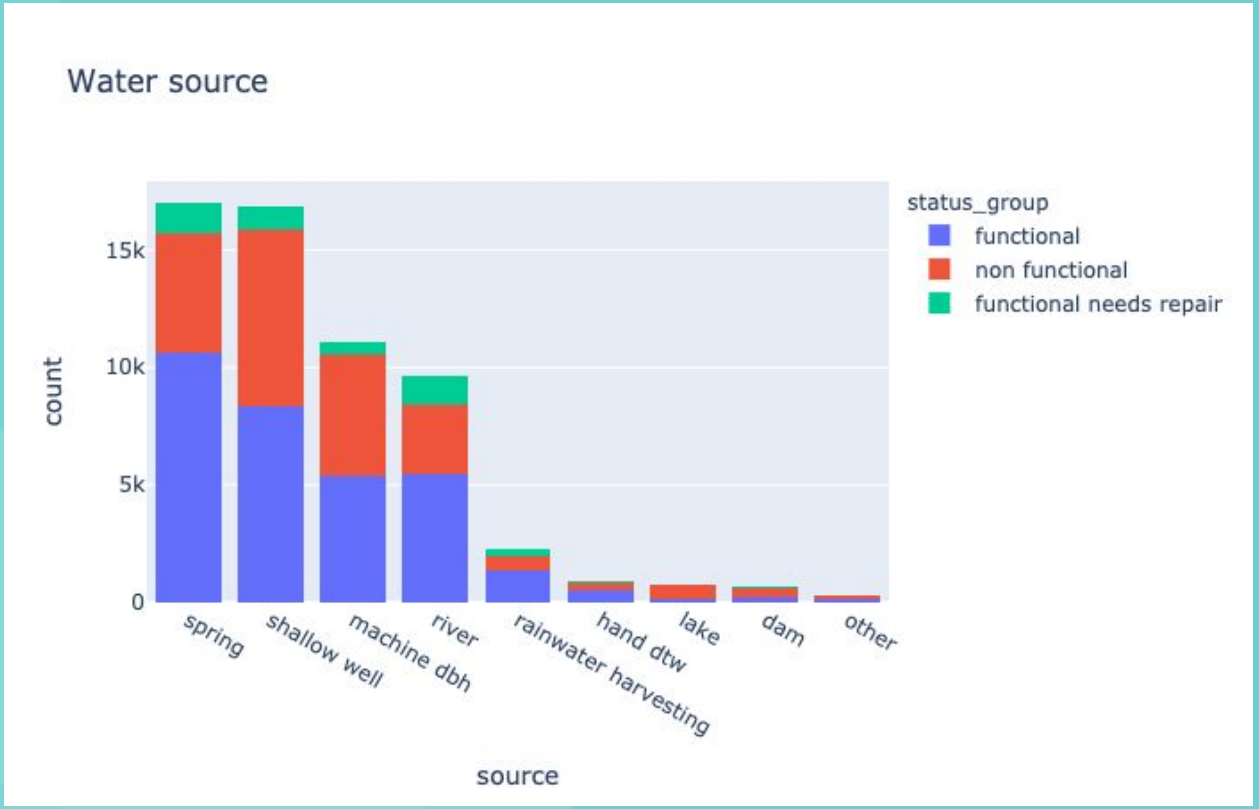
DATA INSIGHTS

SUFFICIENT WATER DEFINES FUNCTIONALITY



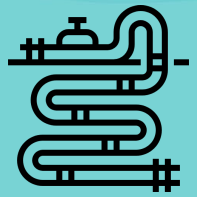
DATA INSIGHTS

WATER SOURCE DEFINES FUNCTIONALITY



PREDICTIVE MODELLING METHODOLOGY

MACHINE LEARNING



PREPROCESSING PIPELINE

LOGISTIC REGRESSION

K-NEAREST NEIGHBORS

DECISION TREE

RANDOM FOREST

ADABOOST

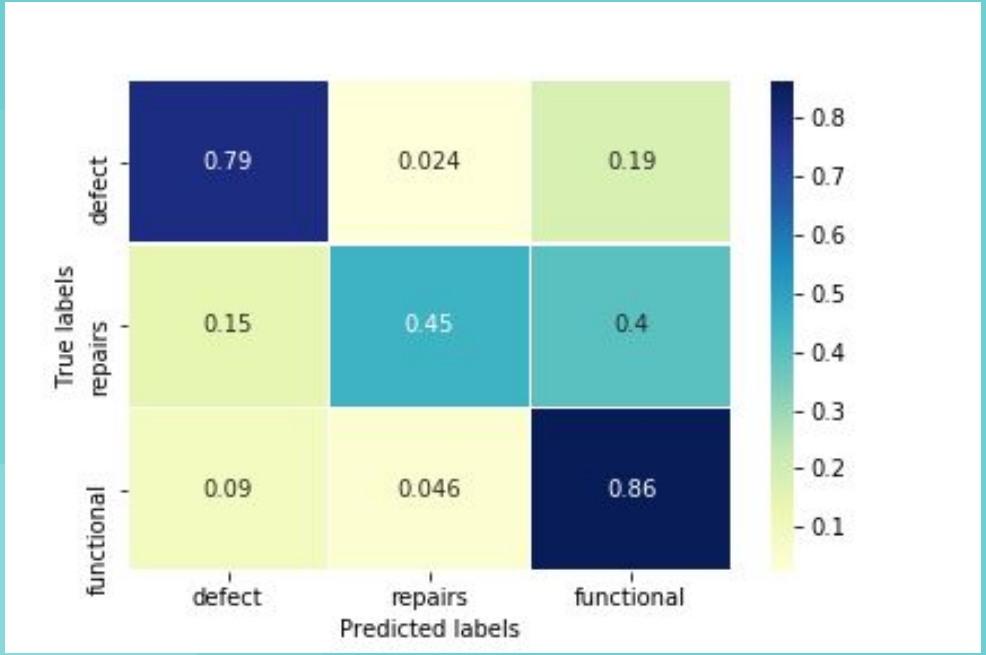
XGBOOST

NEURAL NETWORK

RANDOMIZED GRID SEARCH

MACHINE LEARNING

XGBOOST: BEST MODEL WITH 80 % PRECISION



AND NOW ?

FUTURE WORK

FUNDING OF PROJECTS IN REGIONS WHICH HAVE LESS FUNCTIONAL WELLS

PAY ATTENTION TO PARAMETERS WHICH DETERMINE FUNCTIONALITY OF WELLS

OPTIMIZE IMPUTING STRATEGY OF MISSING VALUES

APPLY/OPTIMIZE NEURAL NETWORK MODEL

CREATE AN APPLICATION

LAST BUT NOT LEAST

ANY QUESTIONS ?

THANK YOU FOR YOUR ATTENTION

github.com/roble-chris

www.linkedin.com/in/christianrobledo90