adammarianacci Updated Notebook and README

08a1a7f · 20 minutes ago History

3698 lines (3698 loc) · 299 KB

Introduction

Final Project Submission

- Student Name: Adam Marianacci
- Student Pace: Flex
- Scheduled project review date/time: TBD
- Instructor Name: Mark Barbour

Business Understanding

It is my job to help the WWFA (Water Wells For Africa) organization identify wells that are in need or repair in Tanzania.

Data Understanding

The data used in this analysis comes from the Taarifa waterpoints dashboard, which aggregates data from the Tanzania Ministry of Water. The final dataframe used in this analysis contained over 38,000 entries. The dataset consisted of various information about waterwells in Tanzania such as the functioning status, water quality, age, source, and altitude to name a few. One limitation of the dataset is that it is a fairly small since we are dealing with predictive modeling. There were also some features that would have been useful but just had too many missing values to use. Another limitation was that many of the features in the dataset were shown to have insignificant importance when it came to predicting wells that were in need of repair. The dataset was suitable for the project because it did reveal some notable features about wells. I was able to gain insight into identifying where repairs were needed to help the WWFA promote access to potable water across Tanzania.

Data Preparation

Data Preperation

In [1]:

```python
# Importing the necessary libraries
import pandas as pd
from datetime import datetime
import numpy as np
import seaborn as sns
import folium
import statsmodels as sm
import sklearn
import sklearn.preprocessing as preprocessing
import matplotlib.pyplot as plt
from scipy import stats
from sklearn import linear_model
from sklearn.linear_model import LogisticRegression
from sklearn.feature_selection import RFE
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.impute import SimpleImputer
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
import warnings
warnings.filterwarnings('ignore')
```

I did not want any information in the dataframe to be truncated. I searched `pandas output truncated` in google and found this solution.

In [2]:

```python
# Set display options to show all rows and columns
pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
```

In [3]:

```python
# Importing the dataframes
```

```python
# Importing the dataframes
df_x = pd.read_csv('data/training_set_values.csv')
df_y = pd.read_csv('data/training_set_labels.csv')
```

In [4]:
```python
# Combining the 2 dataframes into 1 new dataframe
Waterwells_df = pd.concat([df_y, df_x], axis=1)
```

In [5]:
```python
# Previewing the dataframe
Waterwells_df.head()
```

Out[5]:

	id	status_group	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt
0	69572	functional	69572	6000.0	2011-03-14	Roman	1390	Roman	34.938093	-9.856322	
1	8776	functional	8776	0.0	2013-03-06	Grumeti	1399	GRUMETI	34.698766	-2.147466	Z
2	34310	functional	34310	25.0	2013-02-25	Lottery Club	686	World vision	37.460664	-3.821329	M
3	67743	non functional	67743	0.0	2013-01-28	Unicef	263	UNICEF	38.486161	-11.155298	Z Nar
4	19728	functional	19728	0.0	2011-07-13	Action In A	0	Artisan	31.130847	-1.825359	

In [6]:
```python
# Checking the datatypes in my df along with missing values
Waterwells_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59400 entries, 0 to 59399
Data columns (total 42 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   id                     59400 non-null  int64
```

```
 1   status_group           59400 non-null  object 
 2   id                     59400 non-null  int64  
 3   amount_tsh             59400 non-null  float64
 4   date_recorded          59400 non-null  object 
 5   funder                 55765 non-null  object 
 6   gps_height             59400 non-null  int64  
 7   installer              55745 non-null  object 
 8   longitude              59400 non-null  float64
 9   latitude               59400 non-null  float64
 10  wpt_name               59400 non-null  object 
 11  num_private            59400 non-null  int64  
 12  basin                  59400 non-null  object 
 13  subvillage             59029 non-null  object 
 14  region                 59400 non-null  object 
 15  region_code            59400 non-null  int64  
 16  district_code          59400 non-null  int64  
 17  lga                    59400 non-null  object 
 18  ward                   59400 non-null  object 
 19  population             59400 non-null  int64  
 20  public_meeting         56066 non-null  object 
 21  recorded_by            59400 non-null  object 
 22  scheme_management      55523 non-null  object 
 23  scheme_name            31234 non-null  object 
 24  permit                 56344 non-null  object 
 25  construction_year      59400 non-null  int64  
 26  extraction_type        59400 non-null  object 
 27  extraction_type_group  59400 non-null  object 
 28  extraction_type_class  59400 non-null  object 
 29  management             59400 non-null  object 
 30  management_group       59400 non-null  object 
 31  payment                59400 non-null  object 
 32  payment_type           59400 non-null  object 
 33  water_quality          59400 non-null  object 
 34  quality_group          59400 non-null  object 
 35  quantity               59400 non-null  object 
 36  quantity_group         59400 non-null  object 
 37  source                 59400 non-null  object 
 38  source_type            59400 non-null  object 
 39  source_class           59400 non-null  object 
 40  waterpoint_type        59400 non-null  object 
 41  waterpoint_type_group  59400 non-null  object 
dtypes: float64(3), int64(8), object(31)
memory usage: 19.0+ MB
```

Dropping columns that are not directly related to the business problem and also have high cardinality, making them difficult to

Dropping columns that are not directly related to the business problem and also have high cardinality, making them difficult to one hot encode.

In [7]:
```
# Dropping irrelevant columns from the dataframe, also columns with large amounts of missing data
columns_to_drop = [
    'id', 'scheme_management', 'region', 'region_code',
    'payment', 'public_meeting', 'district_code', 'population','amount_tsh',
    'num_private', 'basin', 'latitude', 'longitude',
    'waterpoint_type_group', 'source_class', 'payment_type', 'management_group', 'recorded_by',
    'extraction_type', 'management',
    'source_type', 'extraction_type_group', 'permit', 'funder',
    'date_recorded', 'installer', 'ward', 'scheme_name', 'wpt_name', 'lga', 'subvillage'
]

Waterwells_df = Waterwells_df.drop(columns_to_drop, axis=1, errors='ignore')
```

Setting up my 'y' value to become a binary class. Needs repair -'1' , Does Not need repair - '0'. I wanted to replace 'functional needs repair to read as a '1' for needing repair.

In [8]:
```
# Create a new column 'needs_repair' by merging the two categories
Waterwells_df['needs_repair'] = Waterwells_df['status_group'].replace(
    {'functional': 0, 'non functional': 1,
     'functional needs repair': 1})

# Drop the original 'status_group' column
Waterwells_df.drop('status_group', axis=1, inplace=True)

#Display the updated DataFrame
Waterwells_df.head()
```

Out[8]:

	gps_height	construction_year	extraction_type_class	water_quality	quality_group	quantity	quantity_group	sou
0	1390	1999	gravity	soft	good	enough	enough	sp
1	1399	2010	gravity	soft	good	insufficient	insufficient	rainw harves
2	686	2009	gravity	soft	good	enough	enough	(

3	263	1986	submersible	soft	good	dry	dry	macl
4	0	0	gravity	soft	good	seasonal	seasonal	rainw harves

I wanted to change the construction year into a new column 'age' so it could be easier to work with.

In [9]:
```
#dropping the missing values from the 'construction_year' column and creating a new df
Construction_Year_df = Waterwells_df[Waterwells_df['construction_year'] != 0]

# Calculate the current year
current_year = datetime.now().year

# Create a new column 'age' by subtracting construction year from the current year
Construction_Year_df['age'] = current_year - Waterwells_df['construction_year']
```

In [10]:
```
# deleting the 'construction_year' column since we replaced it with an 'age' column
Construction_Year_df = Construction_Year_df.drop('construction_year', axis=1)
```

We have a class imbalance with the majority of wells not needing repair.

In [11]:
```
# Viewing the value counts of 'needs_repair'
Construction_Year_df['needs_repair'].value_counts()
```

Out[11]:
```
0    21704
1    16987
Name: needs_repair, dtype: int64
```

In [12]:
```
# previewing the new df
Construction_Year_df.head()
```

Out[12]:

	gps_height	extraction_type_class	water_quality	quality_group	quantity	quantity_group	source	waterpoint_typ
0	1390	gravity	soft	good	enough	enough	spring	commun standpip
1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	rainwater	commun

1	1399	gravity	soft	good	insufficient	insufficient	harvesting	standpip
2	686	gravity	soft	good	enough	enough	dam	commun standpip multip
3	263	submersible	soft	good	dry	dry	machine dbh	commun standpip multip
5	0	submersible	salty	salty	enough	enough	other	commun standpip multip

The mean of age is 27.12 and the median is 24 which means the distribuition is slightly skewed to the right. There are a few values on the higher end that are pulling the mean up relative to the median.

In [13]:

```
# Looking at some descriptive statistics of the df
Construction_Year_df.describe()
```

Out[13]:

	gps_height	needs_repair	age
count	38691.000000	38691.000000	38691.000000
mean	1002.367760	0.439043	27.185314
std	618.078669	0.496277	12.472045
min	-63.000000	0.000000	11.000000
25%	372.000000	0.000000	16.000000
50%	1154.000000	0.000000	24.000000
75%	1488.000000	1.000000	37.000000
max	2770.000000	1.000000	64.000000

In [14]:

```
# Checking the
Construction_Year_df['waterpoint_type'].value_counts()
```

Out[14]:

```
communal standpipe            21382
```

```
hand pump                      8759
communal standpipe multiple    4261
other                          3837
improved spring                 367
cattle trough                    80
dam                               5
Name: waterpoint_type, dtype: int64
```

In [15]:

```python
# Checking the data types once again and making sure I no longer have any missing values
Construction_Year_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 38691 entries, 0 to 59399
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   gps_height             38691 non-null  int64
 1   extraction_type_class  38691 non-null  object
 2   water_quality          38691 non-null  object
 3   quality_group          38691 non-null  object
 4   quantity               38691 non-null  object
 5   quantity_group         38691 non-null  object
 6   source                 38691 non-null  object
 7   waterpoint_type        38691 non-null  object
 8   needs_repair           38691 non-null  int64
 9   age                    38691 non-null  int64
dtypes: int64(3), object(7)
memory usage: 3.2+ MB
```

In [16]:

```python
# Defining X and y variables
y = Construction_Year_df["needs_repair"]
X = Construction_Year_df.drop("needs_repair", axis=1)
```

In [17]:

```python
# Performing a train, test, split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state=42)
```

In [18]:

```python
# Looking at the number of missing values in each column
X_train.isna().sum()
```

Out[18]:

```
gps_height             0
```

```
extraction_type_class    0
water_quality            0
quality_group            0
quantity                 0
quantity_group           0
source                   0
waterpoint_type          0
age                      0
dtype: int64
```

In [19]:

```
# Create a list of all the categorical features
cols_to_transform = ['quantity_group', 'waterpoint_type','extraction_type_class',
                     'quality_group', 'source',
                     'water_quality', 'quantity']
# Create a dataframe with the new dummy columns created from the cols_to_transform list
X_train = pd.get_dummies(
    data=X_train, columns=cols_to_transform, drop_first=True, dtype=int)
```

In [20]:

```
# Checking to see if all the data is now numerical - yes.
X_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 30952 entries, 3488 to 24205
Data columns (total 43 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   gps_height                              30952 non-null  int64
 1   age                                     30952 non-null  int64
 2   quantity_group_enough                   30952 non-null  int64
 3   quantity_group_insufficient             30952 non-null  int64
 4   quantity_group_seasonal                 30952 non-null  int64
 5   quantity_group_unknown                  30952 non-null  int64
 6   waterpoint_type_communal standpipe      30952 non-null  int64
 7   waterpoint_type_communal standpipe multiple  30952 non-null  int64
 8   waterpoint_type_dam                     30952 non-null  int64
 9   waterpoint_type_hand pump               30952 non-null  int64
 10  waterpoint_type_improved spring         30952 non-null  int64
 11  waterpoint_type_other                   30952 non-null  int64
 12  extraction_type_class_handpump          30952 non-null  int64
 13  extraction_type_class_motorpump         30952 non-null  int64
 14  extraction_type_class_other             30952 non-null  int64
 15  extraction_type_class_rope pump         30952 non-null  int64
 16  extraction_type_class_submersible       30952 non-null  int64
```

```
 16  extraction_type_class_submersible          30952 non-null  int64
 17  extraction_type_class_wind-powered         30952 non-null  int64
 18  quality_group_fluoride                     30952 non-null  int64
 19  quality_group_good                         30952 non-null  int64
 20  quality_group_milky                        30952 non-null  int64
 21  quality_group_salty                        30952 non-null  int64
 22  quality_group_unknown                      30952 non-null  int64
 23  source_hand dtw                            30952 non-null  int64
 24  source_lake                                30952 non-null  int64
 25  source_machine dbh                         30952 non-null  int64
 26  source_other                               30952 non-null  int64
 27  source_rainwater harvesting                30952 non-null  int64
 28  source_river                               30952 non-null  int64
 29  source_shallow well                        30952 non-null  int64
 30  source_spring                              30952 non-null  int64
 31  source_unknown                             30952 non-null  int64
 32  water_quality_fluoride                     30952 non-null  int64
 33  water_quality_fluoride abandoned           30952 non-null  int64
 34  water_quality_milky                        30952 non-null  int64
 35  water_quality_salty                        30952 non-null  int64
 36  water_quality_salty abandoned              30952 non-null  int64
 37  water_quality_soft                         30952 non-null  int64
 38  water_quality_unknown                      30952 non-null  int64
 39  quantity_enough                            30952 non-null  int64
 40  quantity_insufficient                      30952 non-null  int64
 41  quantity_seasonal                          30952 non-null  int64
 42  quantity_unknown                           30952 non-null  int64
dtypes: int64(43)
memory usage: 10.4 MB
```

In [21]:

```
# previewing my new one hot encoded df
X_train.head()
```

Out[21]:

	gps_height	age	quantity_group_enough	quantity_group_insufficient	quantity_group_seasonal	quantity_group_u
3488	1455	19	0	0	1	
12678	229	17	0	1	0	
37313	1588	14	0	1	0	
20930	1466	17	0	0	1	
3639	1542	34	0	1	0	

Scaling the data of 'gps_height' so that it could be represented appropriately.

In [22]:
```
# Defining the columns to scale
column_to_scale = ['gps_height']

# Initialize the scaler
scaler = MinMaxScaler()

# Fit the scaler on the specified columns and transform the data
X_train[column_to_scale] = scaler.fit_transform(X_train[column_to_scale])
```

In [23]:
```
# Inspecting the data to make sure it was scaled
X_train.head()
```

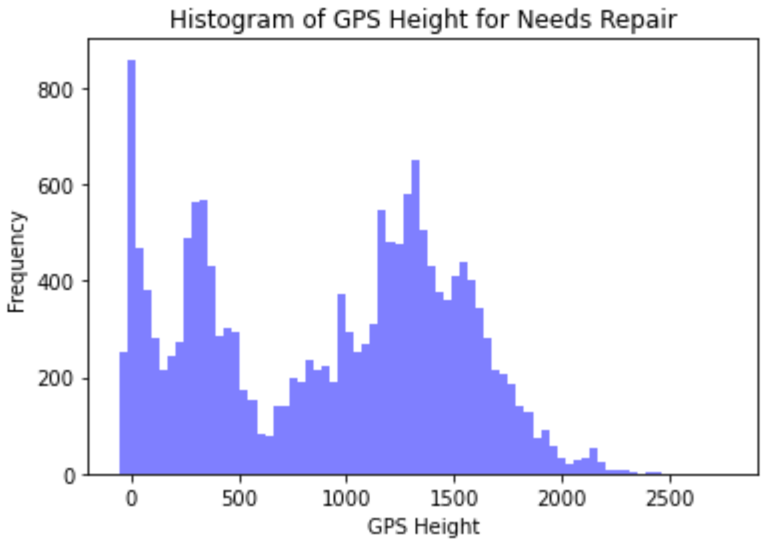
Out[23]:

	gps_height	age	quantity_group_enough	quantity_group_insufficient	quantity_group_seasonal	quantity_group_u
3488	0.535828	19	0	0	1	
12678	0.103071	17	0	1	0	
37313	0.582774	14	0	1	0	
20930	0.539711	17	0	0	1	
3639	0.566537	34	0	1	0	

I wanted to create a visual of how many wells needed repair at different altitudes. The most repairs are needed around sea level. The fewest are needed over 2,000 feet. However this could be due to just fewer wells exist at higher altitudes.

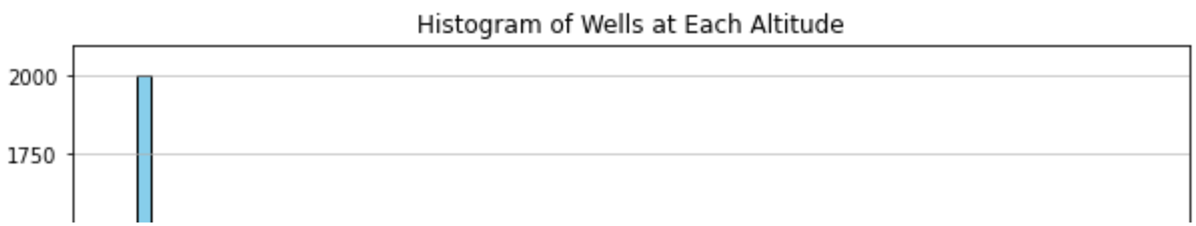
In [24]:
```
# Filtering the data based on 'needs_repair'
needs_repair_histogram = Construction_Year_df[Construction_Year_df['needs_repair'] == 1]['gps_height

#plotting a histogram
plt.hist(needs_repair_histogram, bins=75, color='blue', alpha=0.5)
plt.xlabel('GPS Height')
plt.ylabel('Frequency')
plt.title('Histogram of GPS Height for Needs Repair')
plt.show()
```

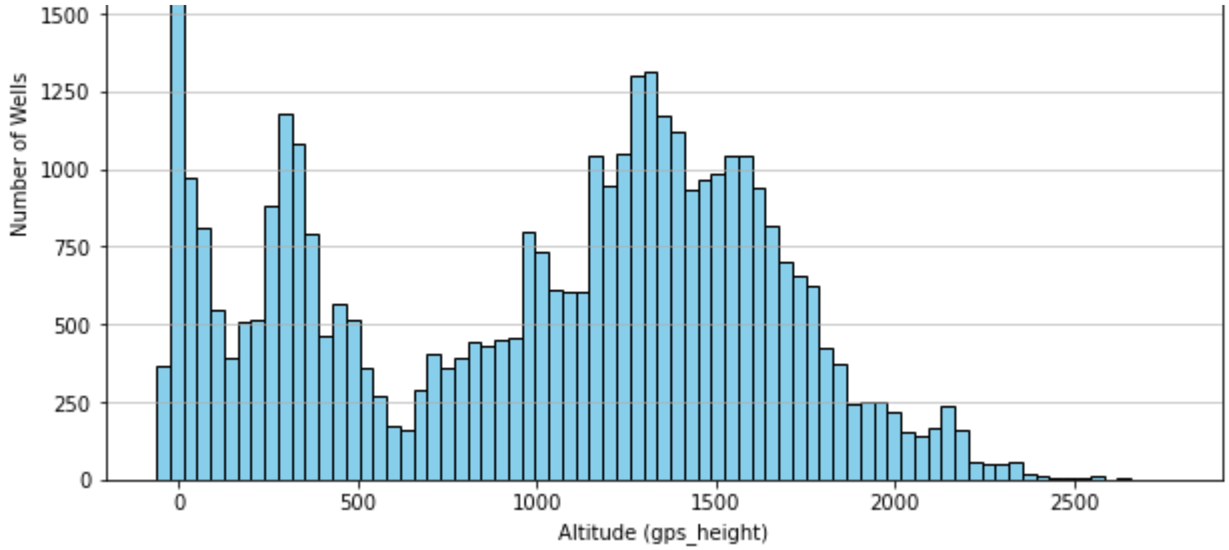
```
plt.show()
```



Next I wanted to see the total number of wells at each altitude. Yes we have the most wells near sea level and the fewest at an altitude of 2300 ft or higher.

In [25]:
```
# Create a histogram
plt.figure(figsize=(10, 6))
plt.hist(Construction_Year_df['gps_height'], bins=75, color='skyblue', edgecolor='black')

# Customize the plot
plt.title('Histogram of Wells at Each Altitude')
plt.xlabel('Altitude (gps_height)')
plt.ylabel('Number of Wells')
plt.grid(axis='y', alpha=0.75)

# Show the plot
plt.show()
```



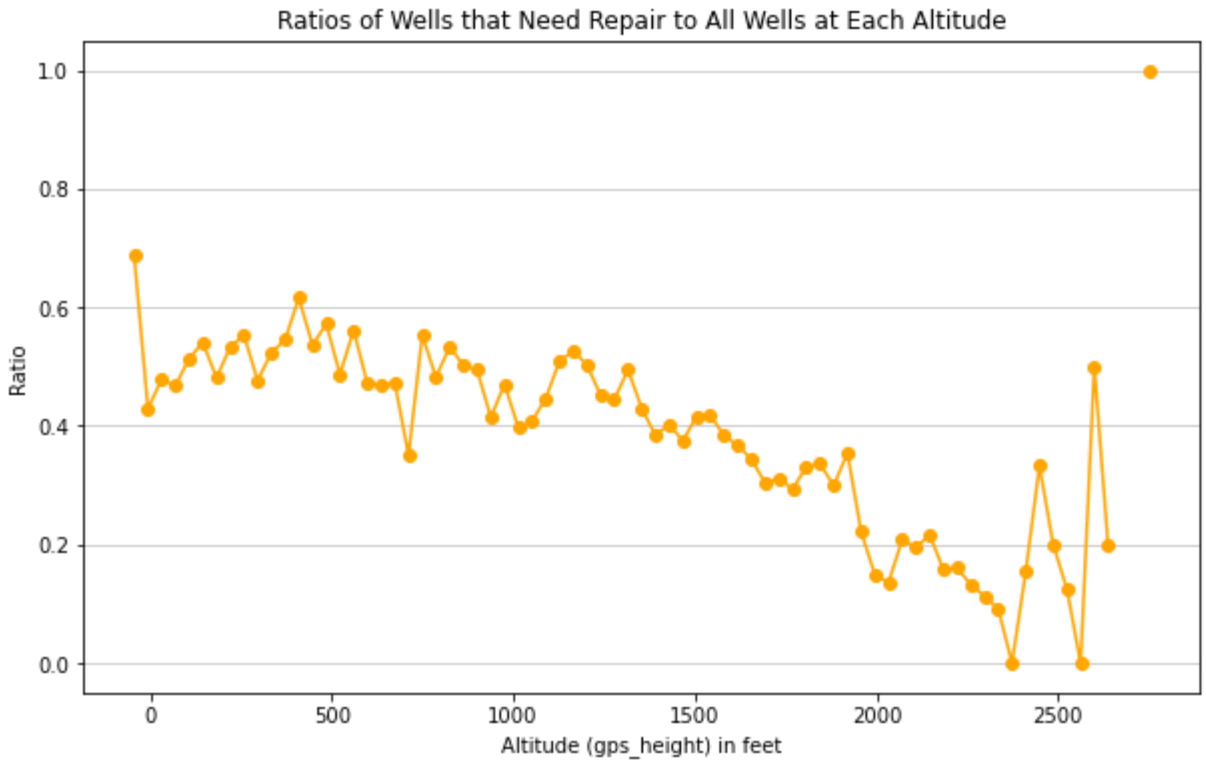


Finally I wanted to create a visual for the ratio of wells that need repair to the total number of wells at each altitude.

In [26]:

```
# Create a histogram for 'gps_height' for all wells
all_histogram, bin_edges_all = np.histogram(Construction_Year_df['gps_height'], bins=75)

# Create a histogram for 'gps_height' for wells that need repair
needs_repair_histogram, bin_edges_needs_repair = np.histogram(
    Construction_Year_df[Construction_Year_df['needs_repair'] == 1]['gps_height'], bins=75)

# Calculate the ratios
ratios = needs_repair_histogram / all_histogram.astype(float)

# Calculate the bin centers
bin_centers = (bin_edges_all[:-1] + bin_edges_all[1:]) / 2

# Plot the ratios
plt.figure(figsize=(10, 6))
plt.plot(bin_centers, ratios, color='orange', marker='o')

# Customize the plot
plt.title('Ratios of Wells that Need Repair to All Wells at Each Altitude')
plt.xlabel('Altitude (gps_height) in feet')
plt.ylabel('Ratio')
plt.grid(axis='y', alpha=0.75)
```

```
# Show the plot
plt.show()
```



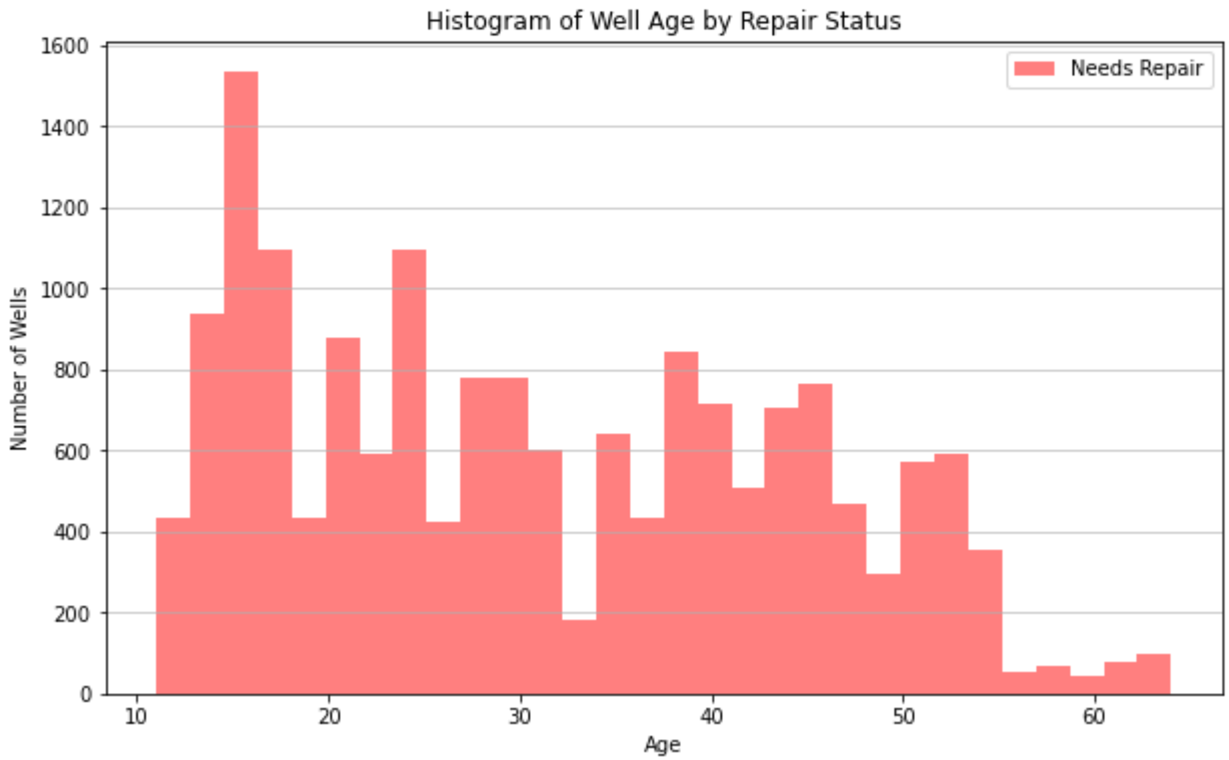
The above graph shows the relationship is generally negative. As altitude increases the repair ratio decreases. However around the 2,400 ft mark the relationship turns generally positive and repair ratio starts to increase.

Next I wanted to get some visuals related to 'age' and 'repairs'.

In [27]:

```
# Filtering data for wells that need repair and those that don't
needs_repair_age = Construction_Year_df[Construction_Year_df['needs_repair'] == 1]['age']

# Create histograms for age of wells
plt.figure(figsize=(10, 6))
plt.hist(needs_repair_age, bins=30, alpha=0.5, color='red', label='Needs Repair')

# Customize the plot
plt.title('Histogram of Well Age by Repair Status')
```

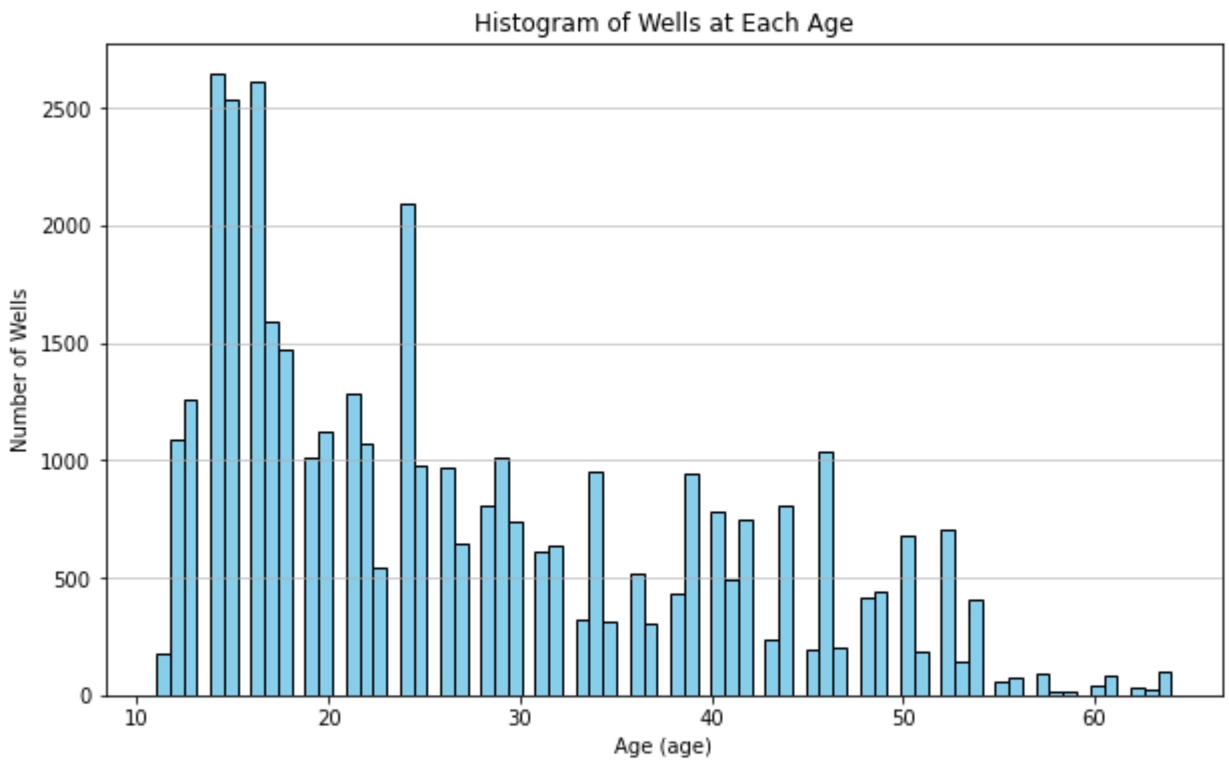
```
plt.xlabel('Age')
plt.ylabel('Number of Wells')
plt.legend()
plt.grid(axis='y', alpha=0.75)

# Show the plot
plt.show()
```



In [28]:

```
# Create a histogram
plt.figure(figsize=(10, 6))
plt.hist(Construction_Year_df['age'], bins=75, color='skyblue', edgecolor='black')

# Customize the plot
plt.title('Histogram of Wells at Each Age')
plt.xlabel('Age (age)')
plt.ylabel('Number of Wells')
plt.grid(axis='y', alpha=0.75)

# Show the plot
```

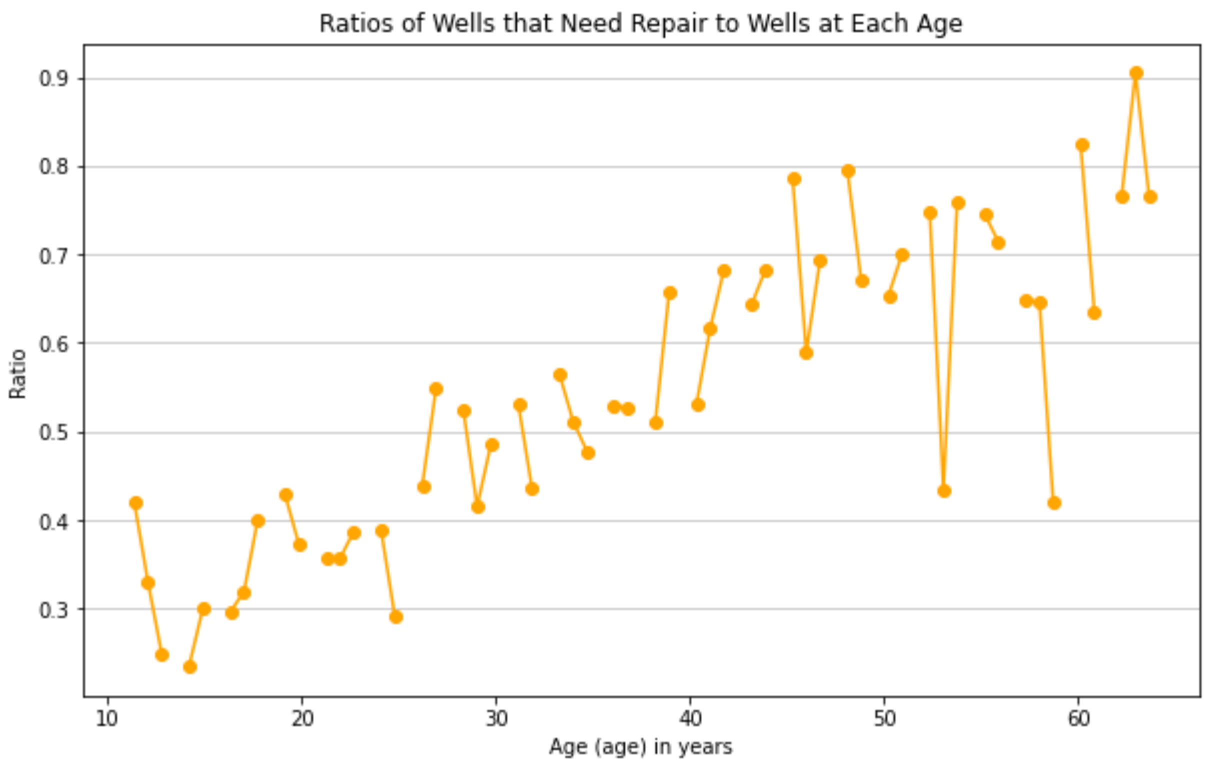
```python
plt.show()
```



I typed `calculating the bin centers in python` into google and found this solution

In [29]:
```python
# Create a histogram for 'age' for all wells
all_histogram_age, bin_edges_all = np.histogram(Construction_Year_df['age'], bins=75)

# Create a histogram for 'gps_height' for wells that need repair
needs_repair_histo, bin_edges_needs_repair = np.histogram(
    Construction_Year_df[Construction_Year_df['needs_repair'] == 1]['age'], bins=75)

# Calculate the ratios
ratios = needs_repair_histo / all_histogram_age.astype(float)

# Calculate the bin centers
bin_centers = (bin_edges_all[:-1] + bin_edges_all[1:]) / 2

# Plot the ratios
```

```
plt.figure(figsize=(10, 6))
plt.plot(bin_centers, ratios, color='orange', marker='o')

# Customize the plot
plt.title('Ratios of Wells that Need Repair to Wells at Each Age')
plt.xlabel('Age (age) in years')
plt.ylabel('Ratio')
plt.grid(axis='y', alpha=0.75)

# Show the plot
plt.show()
```



The above graph shows that there is clearly a positive relationship between the age of a well and the ratio of repairs needed with around the age of 30 roughly 50% of wells are not functioning.

Modeling

In [30]:
```python
# Building a logistic regression model
logreg = LogisticRegression(fit_intercept=False, C=1e12, solver='liblinear')
model_log = logreg.fit(X_train, y_train)
model_log
```

Out[30]:
```
LogisticRegression(C=1000000000000.0, fit_intercept=False, solver='liblinear')
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.

On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

The classifier was about 74% accurate on the training data which is not great.

In [31]:
```python
# Checking the performance on the training data
y_hat_train = logreg.predict(X_train)

train_residuals = np.abs(y_train - y_hat_train)
print(pd.Series(train_residuals, name="Residuals (counts)").value_counts())
print()
print(pd.Series(train_residuals, name="Residuals (proportions)").value_counts(normalize=True))
```

```
0    22982
1     7970
Name: Residuals (counts), dtype: int64

0    0.742505
1    0.257495
Name: Residuals (proportions), dtype: float64
```

In [32]:
```python
# Looking at the number of missing values in each column
X_test.isna().sum()
```

Out[32]:
```
gps_height              0
extraction_type_class   0
water_quality           0
quality_group           0
quantity                0
quantity_group          0
source                  0
waterpoint_type         0
age                     0
dtype: int64
```

dtype: int64

In [33]:
```
# Create a list of all the categorical features
cols_to_transform = ['quantity_group', 'waterpoint_type','extraction_type_class',
                     'quality_group', 'source',
                     'water_quality', 'quantity']
# Create a dataframe with the new dummy columns created from the cols_to_transform list
X_test = pd.get_dummies(
    data=X_test, columns=cols_to_transform, drop_first=True, dtype=int)
```

In [34]:
```
# Fit the scaler on the specified columns and transform the data
X_test[column_to_scale] = scaler.fit_transform(X_test[column_to_scale])
```

In [35]:
```
logreg.score(X_test, y_test)
```

Out[35]: 0.737175345651893

We are still about 74% accuarate on our test data.

In [36]:
```
y_hat_test = logreg.predict(X_test)

test_residuals = np.abs(y_test - y_hat_test)
print(pd.Series(test_residuals, name="Residuals (counts)").value_counts())
print()
print(pd.Series(test_residuals, name="Residuals (proportions)").value_counts(normalize=True))
```

```
0    5705
1    2034
Name: Residuals (counts), dtype: int64

0    0.737175
1    0.262825
Name: Residuals (proportions), dtype: float64
```

The cross validation scores are showing all close to 74% on our 10 folds, showing that we are still consistent with multiple samples from the data.

In [37]:
```
# Getting the cross validation score from our log regression model with X_train and y_train values
cvscore = cross_val_score(logreg, X_train, y_train.values, cv=10)
```

```
cvscore = cross_val_score(logreg, X_train, y_train.values, cv=10)
```

In [38]:
```
# Viewing the scores for the 10 folds we wanted to see, they are all fairly consisten to around 74%
cvscore
```

Out[38]:
```
array([0.74031008, 0.74903101, 0.7450727 , 0.72471729, 0.74087237,
       0.74894992, 0.73893376, 0.74216478, 0.74927302, 0.7457189 ])
```

In [39]:
```
# Confirming the avg cross validation score
np.average(cvscore)
```

Out[39]:
```
0.7425043831636422
```

In [40]:
```
# Looking at standard deviation, this score shows to be very close to the mean
np.std(cvscore)
```

Out[40]:
```
0.006954203732412136
```

Building a single decision tree, this model did not show an improvement from logistic regression. The accuracy which averages precision and recall was at about 72%. It showed gps_height and altitude to be the most important features with gps_height being the most with a score of 0.47 which shows that there is a significant relationship with a well needing repair.

In [41]:
```
# Create the classifier, fit it on the training data and make predictions on the test set
clf = DecisionTreeClassifier(criterion='entropy')

clf.fit(X_train, y_train)
```

Out[41]:
```
DecisionTreeClassifier(criterion='entropy')
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [42]:
```
# Using the trained classifier 'clf'
#to predict the labels for the instances represented by the features in the X_test
#storing the predicted labels into 'y_pred'
y_pred = clf.predict(X_test)
```

In [43]:
```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.75      0.75      0.75      4337
           1       0.68      0.68      0.68      3402

    accuracy                           0.72      7739
   macro avg       0.72      0.72      0.72      7739
weighted avg       0.72      0.72      0.72      7739
```

In [44]:
```
# getting our feature_importance scores
clf.feature_importances_
```

Out[44]:
```
array([4.77296895e-01, 1.75725281e-01, 3.92446650e-02, 4.11335811e-04,
       4.31792838e-02, 0.00000000e+00, 1.89098925e-02, 1.63051465e-02,
       1.76842961e-04, 2.23580502e-03, 1.44581737e-03, 7.86451547e-02,
       3.60422949e-03, 4.50229992e-03, 9.25297256e-03, 2.89595882e-03,
       1.27481799e-02, 4.69687431e-04, 2.50081878e-04, 4.26680561e-03,
       4.61563619e-04, 4.43505715e-03, 2.17328209e-03, 1.94301490e-03,
       1.84158430e-03, 1.39792539e-02, 1.78166082e-03, 6.29810646e-03,
       1.29336265e-02, 7.39046389e-03, 1.54113132e-02, 3.06395615e-04,
       6.92290963e-04, 1.52108630e-04, 3.79485884e-04, 3.84473857e-03,
       1.43231010e-03, 5.38319872e-03, 1.03254095e-03, 3.33255721e-04,
       2.20866454e-02, 3.94461024e-04, 3.74730587e-03])
```

In [45]:
```
# With correlating columns
print("clf.feature_importances_:", clf.feature_importances_)
print("X.columns:", X_train.columns)
```

```
clf.feature_importances_: [4.77296895e-01 1.75725281e-01 3.92446650e-02 4.11335811e-04
 4.31792838e-02 0.00000000e+00 1.89098925e-02 1.63051465e-02
 1.76842961e-04 2.23580502e-03 1.44581737e-03 7.86451547e-02
 3.60422949e-03 4.50229992e-03 9.25297256e-03 2.89595882e-03
 1.27481799e-02 4.69687431e-04 2.50081878e-04 4.26680561e-03
 4.61563619e-04 4.43505715e-03 2.17328209e-03 1.94301490e-03
 1.84158430e-03 1.39792539e-02 1.78166082e-03 6.29810646e-03
 1.29336265e-02 7.39046389e-03 1.54113132e-02 3.06395615e-04
 6.92290963e-04 1.52108630e-04 3.79485884e-04 3.84473857e-03
 1.43231010e-03 5.38319872e-03 1.03254095e-03 3.33255721e-04
 2.20866454e-02 3.94461024e-04 3.74730587e-03]
```

```
X.columns: Index(['gps_height', 'age', 'quantity_group_enough',
       'quantity_group_insufficient', 'quantity_group_seasonal',
       'quantity_group_unknown', 'waterpoint_type_communal standpipe',
       'waterpoint_type_communal standpipe multiple', 'waterpoint_type_dam',
       'waterpoint_type_hand pump', 'waterpoint_type_improved spring',
       'waterpoint_type_other', 'extraction_type_class_handpump',
       'extraction_type_class_motorpump', 'extraction_type_class_other',
       'extraction_type_class_rope pump', 'extraction_type_class_submersible',
       'extraction_type_class_wind-powered', 'quality_group_fluoride',
       'quality_group_good', 'quality_group_milky', 'quality_group_salty',
       'quality_group_unknown', 'source_hand dtw', 'source_lake',
       'source_machine dbh', 'source_other', 'source_rainwater harvesting',
       'source_river', 'source_shallow well', 'source_spring',
       'source_unknown', 'water_quality_fluoride',
       'water_quality_fluoride abandoned', 'water_quality_milky',
       'water_quality_salty', 'water_quality_salty abandoned',
       'water_quality_soft', 'water_quality_unknown', 'quantity_enough',
       'quantity_insufficient', 'quantity_seasonal', 'quantity_unknown'],
      dtype='object')
```

gps_height and age were really the only 2 significant features

In [46]:
```
# Setting up a cleaner way of viewing them in a DF
features = pd.DataFrame(clf.feature_importances_, index=X_train.columns, columns=['Importance'])
print(features)
```

```
                                             Importance
gps_height                                     0.477297
age                                            0.175725
quantity_group_enough                          0.039245
quantity_group_insufficient                    0.000411
quantity_group_seasonal                        0.043179
quantity_group_unknown                         0.000000
waterpoint_type_communal standpipe             0.018910
waterpoint_type_communal standpipe multiple    0.016305
waterpoint_type_dam                            0.000177
waterpoint_type_hand pump                      0.002236
waterpoint_type_improved spring                0.001446
waterpoint_type_other                          0.078645
extraction_type_class_handpump                 0.003604
extraction_type_class_motorpump                0.004502
extraction_type_class_other                    0.009253
extraction_type_class_rope pump                0.002896
```

```
extraction_type_class_submersible              0.012748
extraction_type_class_wind-powered             0.000470
quality_group_fluoride                         0.000250
quality_group_good                             0.004267
quality_group_milky                            0.000462
quality_group_salty                            0.004435
quality_group_unknown                          0.002173
source_hand dtw                                0.001943
source_lake                                    0.001842
source_machine dbh                             0.013979
source_other                                   0.001782
source_rainwater harvesting                    0.006298
source_river                                   0.012934
source_shallow well                            0.007390
source_spring                                  0.015411
source_unknown                                 0.000306
water_quality_fluoride                         0.000692
water_quality_fluoride abandoned               0.000152
water_quality_milky                            0.000379
water_quality_salty                            0.003845
water_quality_salty abandoned                  0.001432
water_quality_soft                             0.005383
water_quality_unknown                          0.001033
quantity_enough                                0.000333
quantity_insufficient                          0.022087
quantity_seasonal                              0.000394
quantity_unknown                               0.003747
```

Building a Random Forest Model. This model improved slightly by showing a 75% on accuracy (f-1 score). This was a slight improvement from our 74% on our baseline logistic regression model but still not great.

In [47]:

```python
#  initializing a Random Forest classifier object that can then be trained on data and used to make 
rf = RandomForestClassifier()
```

In [48]:

```python
# fitting the training and testing data to the model
rf.fit(X_train, y_train)
```

Out[48]:

```
RandomForestClassifier()
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [49]:

```
# Using the trained classifier 'rf'
#to predict the labels for the instances represented by the features in the X_test
#storing the predicted labels into 'y_pred' and 'y_train_pred' for X_train
y_pred = rf.predict(X_test)
y_train_pred = rf.predict(X_train)
```

In [50]:

```
# Checking the accuracy of the model
rf.score(X_test, y_test)
```

Out[50]: 0.7552655381832278

In [51]:

```
# Viewing the classification report for y_test and y_pred
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.77      0.80      0.78      4337
           1       0.73      0.70      0.72      3402

    accuracy                           0.76      7739
   macro avg       0.75      0.75      0.75      7739
weighted avg       0.75      0.76      0.75      7739
```

In [52]:

```
# Viewing the classification report for y_train, y_train_pred
print(classification_report(y_train, y_train_pred))
```

```
              precision    recall  f1-score   support

           0       0.98      0.99      0.99     17367
           1       0.99      0.97      0.98     13585

    accuracy                           0.98     30952
   macro avg       0.98      0.98      0.98     30952
weighted avg       0.98      0.98      0.98     30952
```

The training data is performing much better than the testing data which means the model is most likely overfitting.

Again, gps_height and age were the only 2 significant features

In [53]:
```python
# Checking to see what features were the most important in the model
features = pd.DataFrame(rf.feature_importances_, index = X_train.columns)
print(features)
```

```
                                                0
gps_height                               0.463368
age                                      0.210221
quantity_group_enough                    0.029207
quantity_group_insufficient              0.016585
quantity_group_seasonal                  0.013141
quantity_group_unknown                   0.001329
waterpoint_type_communal standpipe       0.022509
waterpoint_type_communal standpipe multiple  0.013124
waterpoint_type_dam                      0.000113
waterpoint_type_hand pump                0.007390
waterpoint_type_improved spring          0.001992
waterpoint_type_other                    0.041896
extraction_type_class_handpump           0.008081
extraction_type_class_motorpump          0.004178
extraction_type_class_other              0.031825
extraction_type_class_rope pump          0.001923
extraction_type_class_submersible        0.008738
extraction_type_class_wind-powered       0.000485
quality_group_fluoride                   0.000595
quality_group_good                       0.003724
quality_group_milky                      0.000300
quality_group_salty                      0.002098
quality_group_unknown                    0.006713
source_hand dtw                          0.001365
source_lake                              0.004826
source_machine dbh                       0.007304
source_other                             0.001555
source_rainwater harvesting              0.004618
source_river                             0.006274
source_shallow well                      0.005932
source_spring                            0.009369
source_unknown                           0.000294
water_quality_fluoride                   0.000633
water_quality_fluoride abandoned         0.000124
water_quality_milky                      0.000298
water_quality_salty                      0.002011
water_quality_salty abandoned            0.000758
```

```
water_quality_soft                             0.003436
water_quality_unknown                          0.004200
quantity_enough                                0.026731
quantity_insufficient                          0.017105
quantity_seasonal                              0.012292
quantity_unknown                               0.001343
```

In [54]:

```python
# Sorting the features by most influential to least
features_sorted = features.sort_values(by=0, ascending=False)
print(features_sorted)
```

```
                                                      0
gps_height                                     0.463368
age                                            0.210221
waterpoint_type_other                          0.041896
extraction_type_class_other                    0.031825
quantity_group_enough                          0.029207
quantity_enough                                0.026731
waterpoint_type_communal standpipe             0.022509
quantity_insufficient                          0.017105
quantity_group_insufficient                    0.016585
quantity_group_seasonal                        0.013141
waterpoint_type_communal standpipe multiple    0.013124
quantity_seasonal                              0.012292
source_spring                                  0.009369
extraction_type_class_submersible              0.008738
extraction_type_class_handpump                 0.008081
waterpoint_type_hand pump                      0.007390
source_machine dbh                             0.007304
quality_group_unknown                          0.006713
source_river                                   0.006274
source_shallow well                            0.005932
source_lake                                    0.004826
source_rainwater harvesting                    0.004618
water_quality_unknown                          0.004200
extraction_type_class_motorpump                0.004178
quality_group_good                             0.003724
water_quality_soft                             0.003436
quality_group_salty                            0.002098
water_quality_salty                            0.002011
waterpoint_type_improved spring                0.001992
extraction_type_class_rope pump                0.001923
source_other                                   0.001555
source_hand dtw                                0.001365
```

```
quantity_unknown                            0.001343
quantity_group_unknown                      0.001329
water_quality_salty abandoned               0.000758
water_quality_fluoride                      0.000633
quality_group_fluoride                      0.000595
extraction_type_class_wind-powered          0.000485
quality_group_milky                         0.000300
water_quality_milky                         0.000298
source_unknown                              0.000294
water_quality_fluoride abandoned            0.000124
waterpoint_type_dam                         0.000113
```

Building a second Random Forest model with hyperparameters. This showed to improve the model to about a 78% accuracy. It also showed a 76% on the weighted avg. for recall. I chose to look at the macro avg. to be more conservative as this gave a lower score than the weighted avg.

In [55]:

```python
# Using hyperparameters to hopefully improve the model.
# Adding more trees to the forest to increase performance.
# Using min_samples_split to help control overfitting
# Using max depth so trees can grow deeper and learn more information.
# Using a random state so results will be reproducible across multiple runs.
rf2 = RandomForestClassifier(n_estimators = 1000,
                             criterion = 'entropy',
                             min_samples_split = 10,
                             max_depth = 15,
                             random_state = 42
)
```

In [56]:

```python
# fitting the training and testing data to the model
rf2.fit(X_train, y_train)
```

Out[56]:

```
RandomForestClassifier(criterion='entropy', max_depth=15, min_samples_split=10,
                       n_estimators=1000, random_state=42)
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

This model received a mean accuracy score of 77% which is an improvement.

In [57]:

```python
# Checking the accuracy of the model
```

```
# checking the accuracy of the model
rf2.score(X_test, y_test)
```

Out[57]: 0.7771029848817677

In [58]:

```
# Using the trained classifier 'rf2'
#to predict the labels for the instances represented by the features in the X_test
#storing the predicted labels into 'y_pred2'
y_pred2 = rf2.predict(X_test)
y_train_pred2 = rf2.predict(X_train)
```

In [59]:

```
# Viewing the classification report
print(classification_report(y_test, y_pred2))
```

```
              precision    recall  f1-score   support

           0       0.74      0.92      0.82      4337
           1       0.85      0.60      0.70      3402

    accuracy                           0.78      7739
   macro avg       0.80      0.76      0.76      7739
weighted avg       0.79      0.78      0.77      7739
```

In [60]:

```
# Viewing the classification report for y_test, y_train_pred2)
print(classification_report(y_train, y_train_pred2))
```

```
              precision    recall  f1-score   support

           0       0.78      0.95      0.86     17367
           1       0.91      0.66      0.76     13585

    accuracy                           0.82     30952
   macro avg       0.84      0.80      0.81     30952
weighted avg       0.84      0.82      0.81     30952
```

The training data is still performing better than our testing data, but we have improved the model by getting the scores closer to each other and reduced overfitting. The accuracy is 82% on our training data and 78% on our testing data. The macro avg. of recall is 80% on our training data and 76% on our testing data.

In [61]:

```python
# Checking to see what features were the most important in the model
features = pd.DataFrame(rf2.feature_importances_, index = X_train.columns)
print(features)
```

```
                                                    0
gps_height                                   0.174928
age                                          0.204496
quantity_group_enough                        0.052745
quantity_group_insufficient                  0.030399
quantity_group_seasonal                      0.026672
quantity_group_unknown                       0.003800
waterpoint_type_communal standpipe           0.041878
waterpoint_type_communal standpipe multiple  0.024623
waterpoint_type_dam                          0.000199
waterpoint_type_hand pump                    0.013010
waterpoint_type_improved spring              0.004313
waterpoint_type_other                        0.085681
extraction_type_class_handpump               0.013270
extraction_type_class_motorpump              0.007218
extraction_type_class_other                  0.060175
extraction_type_class_rope pump              0.003420
extraction_type_class_submersible            0.013185
extraction_type_class_wind-powered           0.000707
quality_group_fluoride                       0.001126
quality_group_good                           0.007083
quality_group_milky                          0.000550
quality_group_salty                          0.003862
quality_group_unknown                        0.010975
source_hand dtw                              0.002364
source_lake                                  0.009164
source_machine dbh                           0.012076
source_other                                 0.003458
source_rainwater harvesting                  0.008757
source_river                                 0.009067
source_shallow well                          0.011366
source_spring                                0.018249
source_unknown                               0.000417
water_quality_fluoride                       0.001182
water_quality_fluoride abandoned             0.000174
water_quality_milky                          0.000564
water_quality_salty                          0.003575
water_quality_salty abandoned                0.001344
water_quality_soft                           0.007038
```

```
water_quality_unknown                      0.011520
quantity_enough                            0.051274
quantity_insufficient                      0.032429
quantity_seasonal                          0.028103
quantity_unknown                           0.003561
```

Age and gps_height once again stood out as the 2 features that showed the most importance, this time with age being at the top.

In [62]:

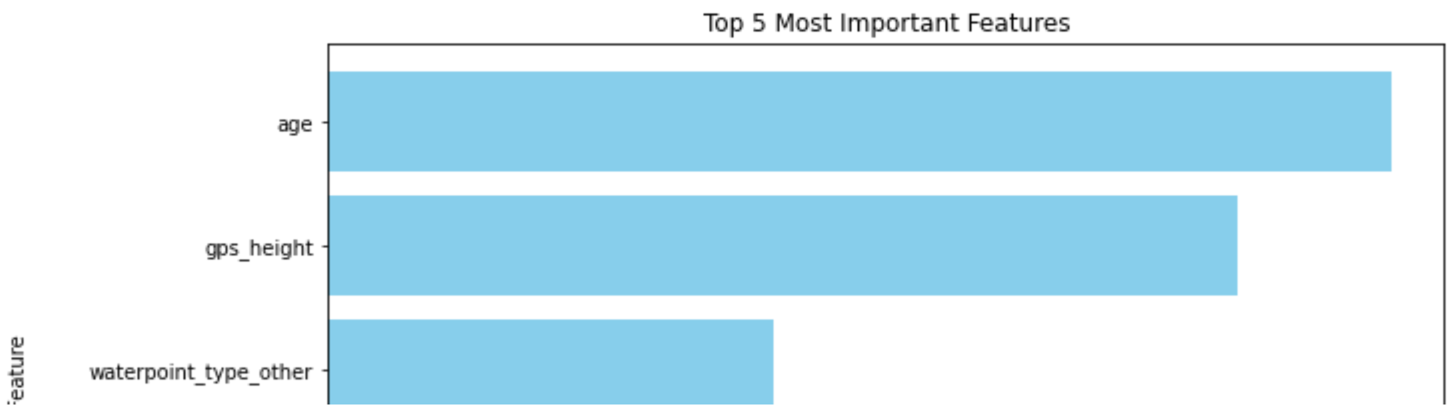
```
# Sorting the features by most influential to least
features_sorted = features.sort_values(by=0, ascending=False)
print(features_sorted)
```

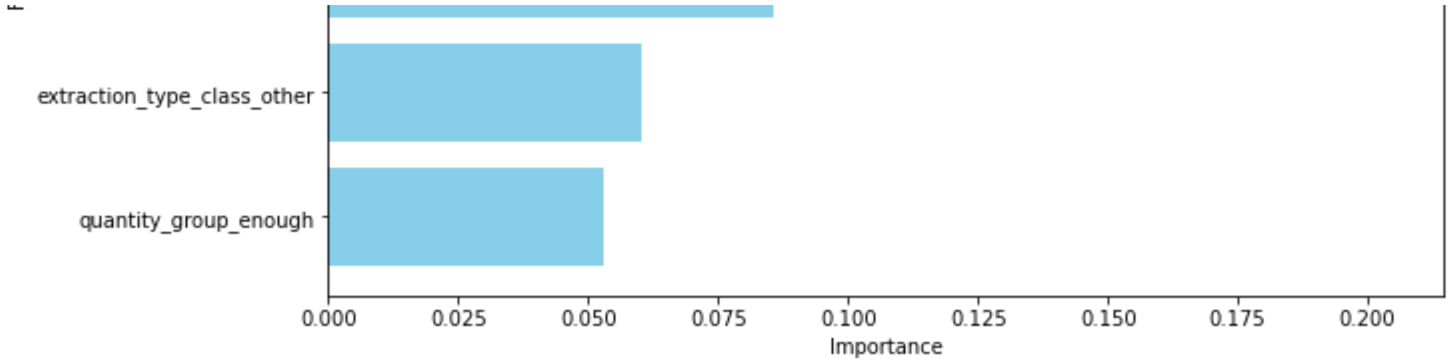
```
                                                  0
age                                        0.204496
gps_height                                 0.174928
waterpoint_type_other                      0.085681
extraction_type_class_other                0.060175
quantity_group_enough                      0.052745
quantity_enough                            0.051274
waterpoint_type_communal standpipe         0.041878
quantity_insufficient                      0.032429
quantity_group_insufficient                0.030399
quantity_seasonal                          0.028103
quantity_group_seasonal                    0.026672
waterpoint_type_communal standpipe multiple  0.024623
source_spring                              0.018249
extraction_type_class_handpump             0.013270
extraction_type_class_submersible          0.013185
waterpoint_type_hand pump                  0.013010
source_machine dbh                         0.012076
water_quality_unknown                      0.011520
source_shallow well                        0.011366
quality_group_unknown                      0.010975
source_lake                                0.009164
source_river                               0.009067
source_rainwater harvesting                0.008757
extraction_type_class_motorpump            0.007218
quality_group_good                         0.007083
water_quality_soft                         0.007038
waterpoint_type_improved spring            0.004313
quality_group_salty                        0.003862
quantity_group_unknown                     0.003800
```

```
water_quality_salty                          0.003575
quantity_unknown                             0.003561
source_other                                 0.003458
extraction_type_class_rope pump              0.003420
source_hand dtw                              0.002364
water_quality_salty abandoned                0.001344
water_quality_fluoride                       0.001182
quality_group_fluoride                       0.001126
extraction_type_class_wind-powered           0.000707
water_quality_milky                          0.000564
quality_group_milky                          0.000550
source_unknown                               0.000417
waterpoint_type_dam                          0.000199
water_quality_fluoride abandoned             0.000174
```

In [63]:

```python
# Selecting the top features
top_features = features_sorted.iloc[:5]  # Selecting the top 5 features

# Extracting feature names and their importance values
feature_names = top_features.index
importance_values = top_features[0]

# Plotting the bar chart
plt.figure(figsize=(10, 6))
plt.barh(feature_names, importance_values, color='skyblue')
plt.xlabel('Importance')
plt.ylabel('Feature')
plt.title('Top 5 Most Important Features')
plt.gca().invert_yaxis()  # Invert y-axis to have the highest importance at the top
plt.show()
```



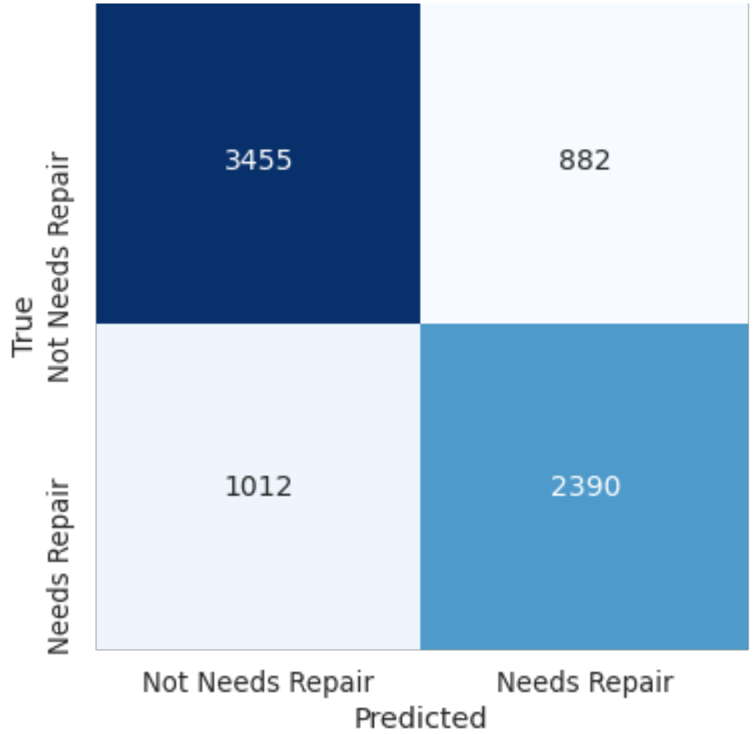


In [64]:
```
# Checking the dimensions of the confusion matrix
print(confusion_matrix(y_test, y_pred))
```

```
[[3455  882]
 [1012 2390]]
```

The confusion matrix shows that our True/Positives are 2,388, our True/Negatives are 3,440. The False/Positives are at 897, and the False/Negatives are 1,014. This sample shows that the model is predicting a FN 13% of the time which is not good.

In [65]:
```
# Generating a confusion matrix
cm = confusion_matrix(y_test, y_pred)

# Set up a figure and axis
plt.figure(figsize=(8, 6))
sns.set(font_scale=1.2)  # Adjust font size for better readability

# Create a heatmap of the confusion matrix
sns.heatmap(cm, annot=True, fmt='g', cmap='Blues', cbar=False,
            annot_kws={"size": 14}, square=True,
            xticklabels=['Not Needs Repair', 'Needs Repair'],
            yticklabels=['Not Needs Repair', 'Needs Repair'])

# Labeling and viewing the cm
plt.xlabel('Predicted')
plt.ylabel('True')
plt.title('Confusion Matrix')
plt.show()
```



Evaluation

My best performing model was my rf2 model which was the second Random Forest model with hyperparameters. It showed a 76% on the macro avg. (where all classes equally contribute to the final averaged metric) of recall. Although this isn't great, it does help in identifying wells that are in need of repair. I focused on recall because it explains how many of the actual positive cases we were able to predict correctly. The confusion matrix showed that the model was falsely identifying wells 13% of the time on a sample size that was 20% of our total data. When it came to the problem of the business understanding it was more of a concern to identify false negatives , labeling wells as not needing repair that are actually in need of repair will lead to people not having access to clean water. It showed age and gps_height as the 2 most important features with "age" as the most important feature which was different from the other models that showed gps_height as the feature of most importance.

Conclusion

The 'rf2' which was the 2nd Random Forest Model with hyperparameters was our best peforming model which showed a 76% macro avg. on recall. Although this wasn't a stellar score in helped to gain insights on wells that should be repaired. We need to gather more data (hundreds of thousands more entries) from features that show higher importance percentages, this will improve the predictive capabilites of our models. I found that there was a positive relationship between the ratio of wells needing repair and the age of a well. I also discovered there is generally a negative relationship between the ratio of wells needing repair and the altitude of a well from slightly below sea level to roughly 2,400 feet above sea level. I noticed after 2,400 feet the relationship changes to a positive one. More analysis needs to be conducted to draw conclusions about this relationship.

Recommendations

I recommend that there should be an age threshold on waterwells that require repair/replacement of every well by the age of 20. My analysis indicates that roughly 50% of wells are in need of repairs by the age of 30. If we send repair specialists to wells starting at the age of 20 we can tackle problems before they become larger issues potenitally leaving people without clean drinking water. I also recommend we gather more data regarding population around the well. Anything mechanical undergoes 'wear and tear' the more it is used. Gathering more information on the population around the wells will show what kind of impact this has on the ratio of wells needing repair. This may also help us understand the relationship of the ratio of wells needing repairs at each altitude, since the reasons were inconclusive. Lastly I recommend gathering more data on geographic location to see what wells were not functioning because of mechanical issues and which wells were not functioning due to a lack of water supply, looking at areas susceptile to droughts would be one example of how further data would be useful to locate problem wells due to geographic location.

Limitations

The main limitation of this dataset was that there were not many features that showed significant importance in our models. There was also a lot of missing values in the dataset, too many to the point where certain features could not be used. Also the final dataframe used consisted of only 38,000 entries, gathering 10x more data on features with greater importance to our target variable will improve our model.

Next Steps

We need to start making repairs mandatory and start replacing wells at the age of 20. We need to look at data regarding population around the well to see if this is having an impact on the lifespan of a well. The more use the well undergoes the quicker it is likely to breakdown I suspect. Having access to this information would certainly help our model. We also need to gather more geographic data around the wells to learn more about the reasons wells are not functioning (mechanical or geographic issues (a drought etc. causing a lack of water supply). Lastly I would like to gather data on how the well is maintained. How frequently are the wells checked to be working properly and by who? trained or untrained people? This could also have an impact on the longevity of a well. Are wells in cities looked after more than ones in rural areas? This would help in locating problem areas for repairs.