Identifying Abnormalities in Energy Consumption

Luke Bravo | Ram Kapistalam
India Lindsay | Mauricio Morales | Jocelyne Walker

OUR STRUCTURE

01

PROBLEM & DATA

Background, data collection, and pre-processing

02

PROCESS

Chosen models, feature selection, and validation

03

FINDINGS

Lessons learned, evaluation, and anomaly detection

BACKGROUND

1

Steadily increasing energy consumption over the past decade

2

Faulty construction, malfunctioning equipment lead to anomalous usage

3

Primitive methods are currently used for fault detection

The opportunity:

Commercial buildings **waste 15-30% of energy** used due to improperly controlled equipment.

A two-part challenge

PREDICTING ENERGY CONSUMPTION

Build a supervised model to identify the **typical usage level** given time series lagged demand and other relevant predictors

IDENTIFYING USAGE ANOMALIES

Develop classification metric to identify **abnormal energy usage** and interpret over- or under-consumption

PROVIDED DATA

Consumption

Timestamp and energy values in Watt-hours for each meter at 10, 15, or 30 minute intervals

Weather

Local temperatures at weather stations near energy meters

Holiday

Timestamps for public holidays where consumption may be lower than expected

Metadata

Site location along with a description of the environment of the meter

Schneider Electric hosted competition and provided data via datadriven.org

PRE-PROCESSING STEPS

1. Merged datasets and selected site 38 with laboratory meter 10
2. Handled null values by removing empty rows
3. Unified energy measurements into kwH
4. Constructed target variable by differencing energy values
5. Feature engineered measurement lags with 1 and 2 differences
6. Created daily and monthly lags to measure seasonality

Our reasoning for Feature Engineering

Time series data is autocorrelated

Created features: previous change in energy usage and twice previous change in energy usage

Consumption varies over time of day and month

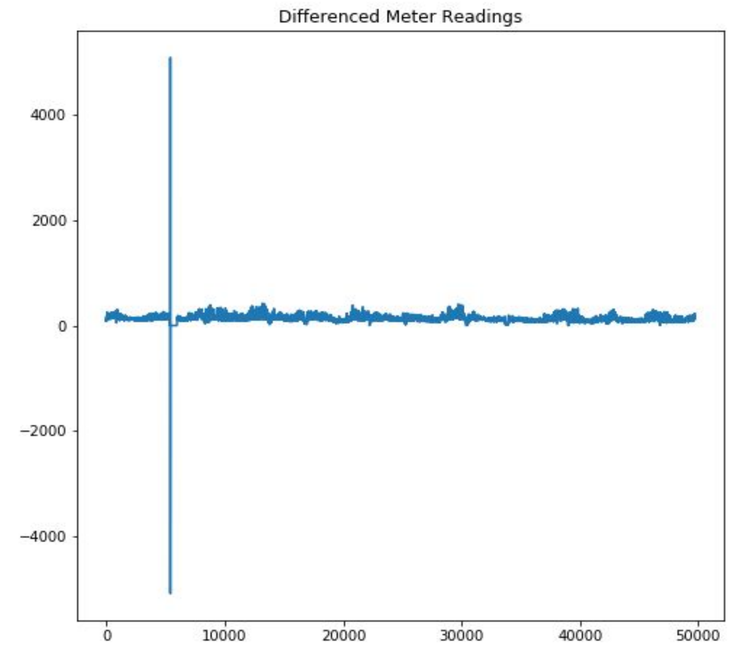
Created features that lagged to the same hour the previous time of day and previous month

Holidays may lead to lower energy consumption

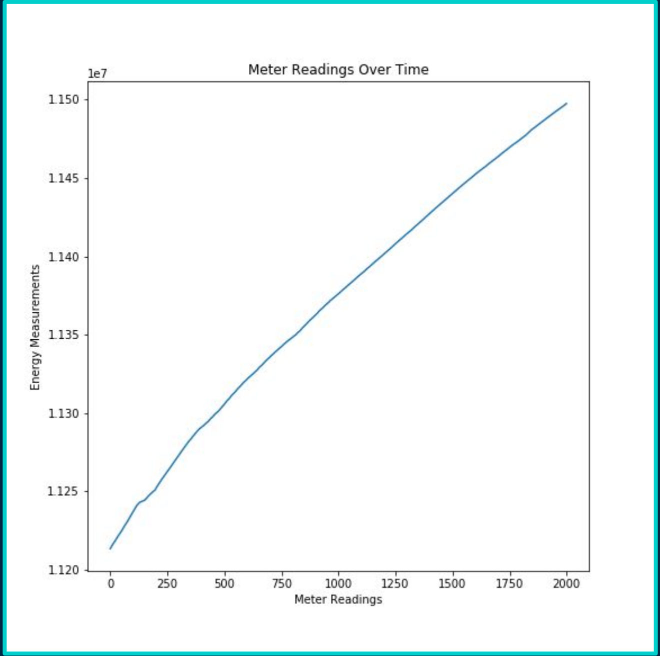
Created binary dummy variable to denote holidays

Outlier detection when examining differences led us to remove some timestamps

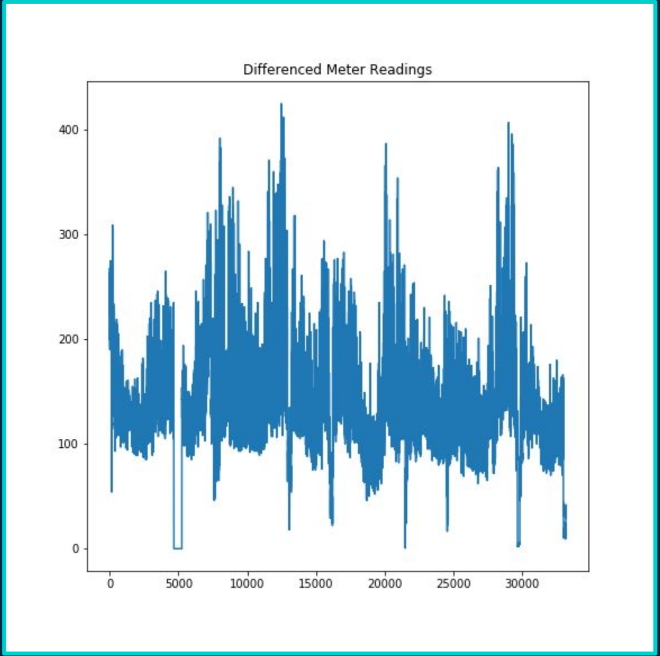
	Timestamp	Values	tv delta
5404	2012-08-20 06:00:00+00:00	11911557.0	0.0
5405	2012-08-20 07:00:00+00:00	11911557.0	0.0
5406	2012-08-20 08:00:00+00:00	11911557.0	0.0
5407	2012-08-20 09:00:00+00:00	11916652.0	5095.0
5408	2012-08-20 10:00:00+00:00	11911557.0	-5095.0
5409	2012-08-20 11:00:00+00:00	11911557.0	0.0
5410	2012-08-20 12:00:00+00:00	11911557.0	0.0



Differencing the Target Variable



Cumulative Energy Measurements



OUR PREDICTOR VARIABLE:

Change in Energy Consumption

OUR VARIABLES

Temperature

Holiday indicator

Additional predictors

Daily Shift

Monthly Shift

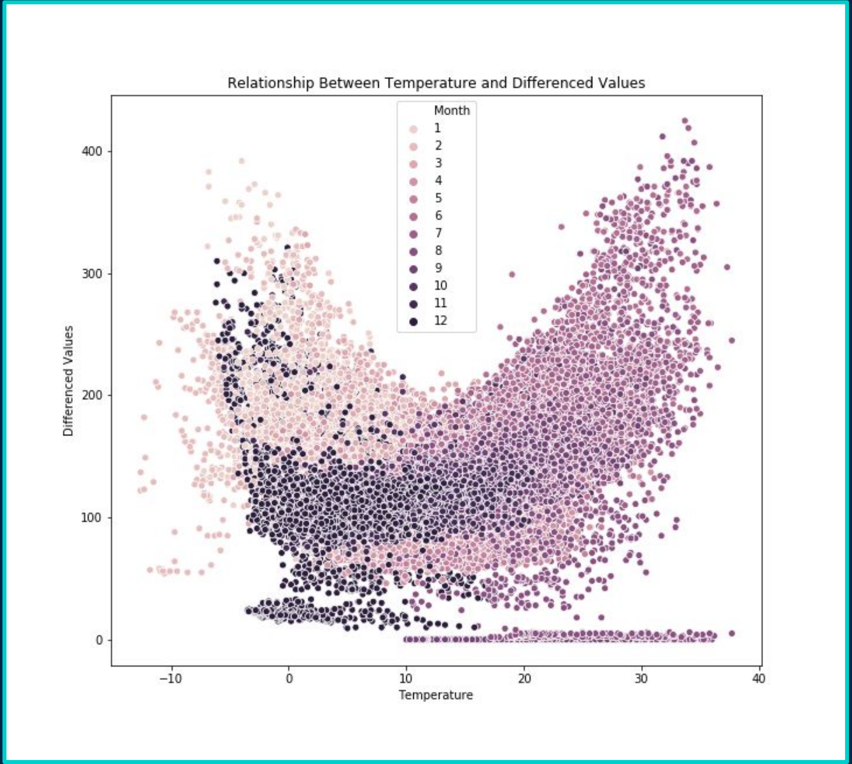
Differenced Measurements

Twice Differenced Measurements

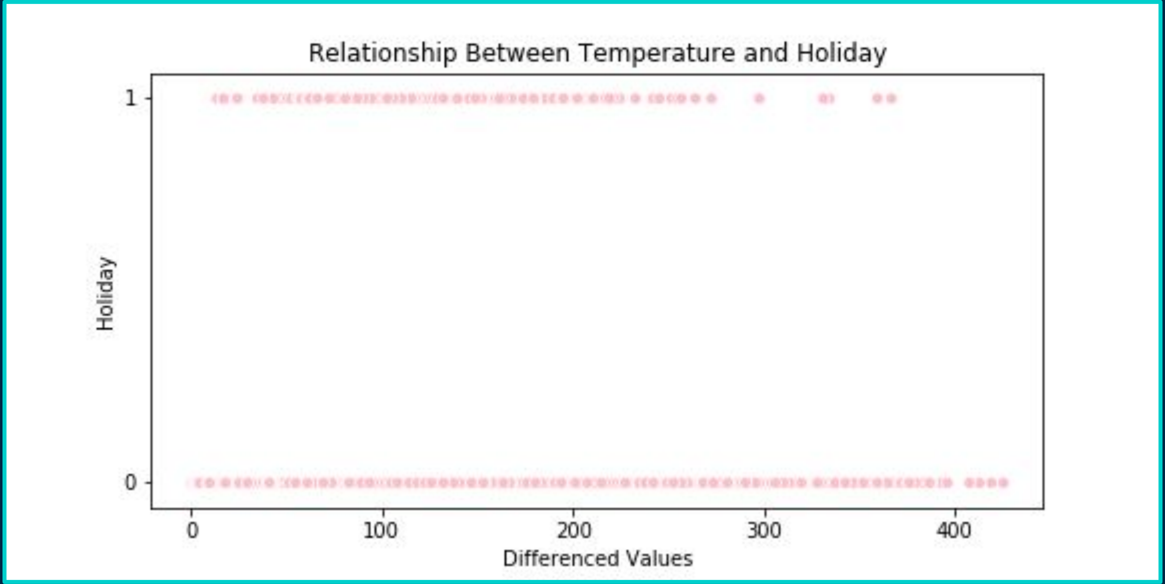
Time series lagged variables

Future Change in Energy Consumption

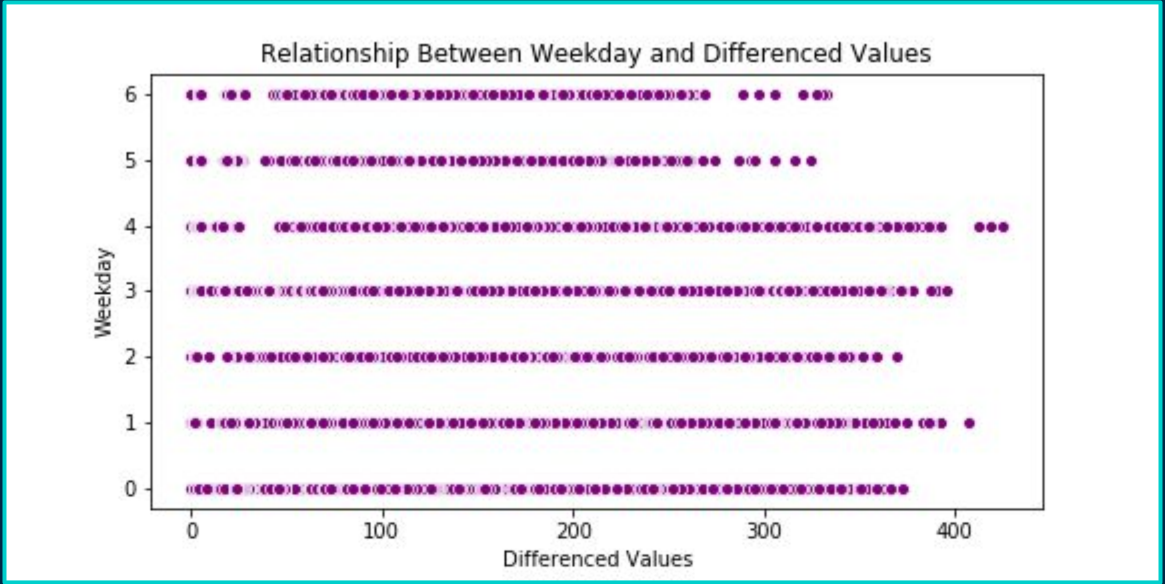
Seasonal relationship between temperature and energy consumption



On average, lower energy consumption during holidays



Relationship Between Day of Week and Energy Consumption



OUR MODELING PROCESS

Train: 2012-2015
Test: 2016-2017

Building distinct models for weekdays and weekends did **not** improve model performance.

Random Forest

Tuned on number of estimators, min samples leaf value, and max features

XGBoost

Trained with max depth of 2 and learning rate of 0.1

Neural Net

Tuned on hidden layer size and batch size, with negative RMSE scoring

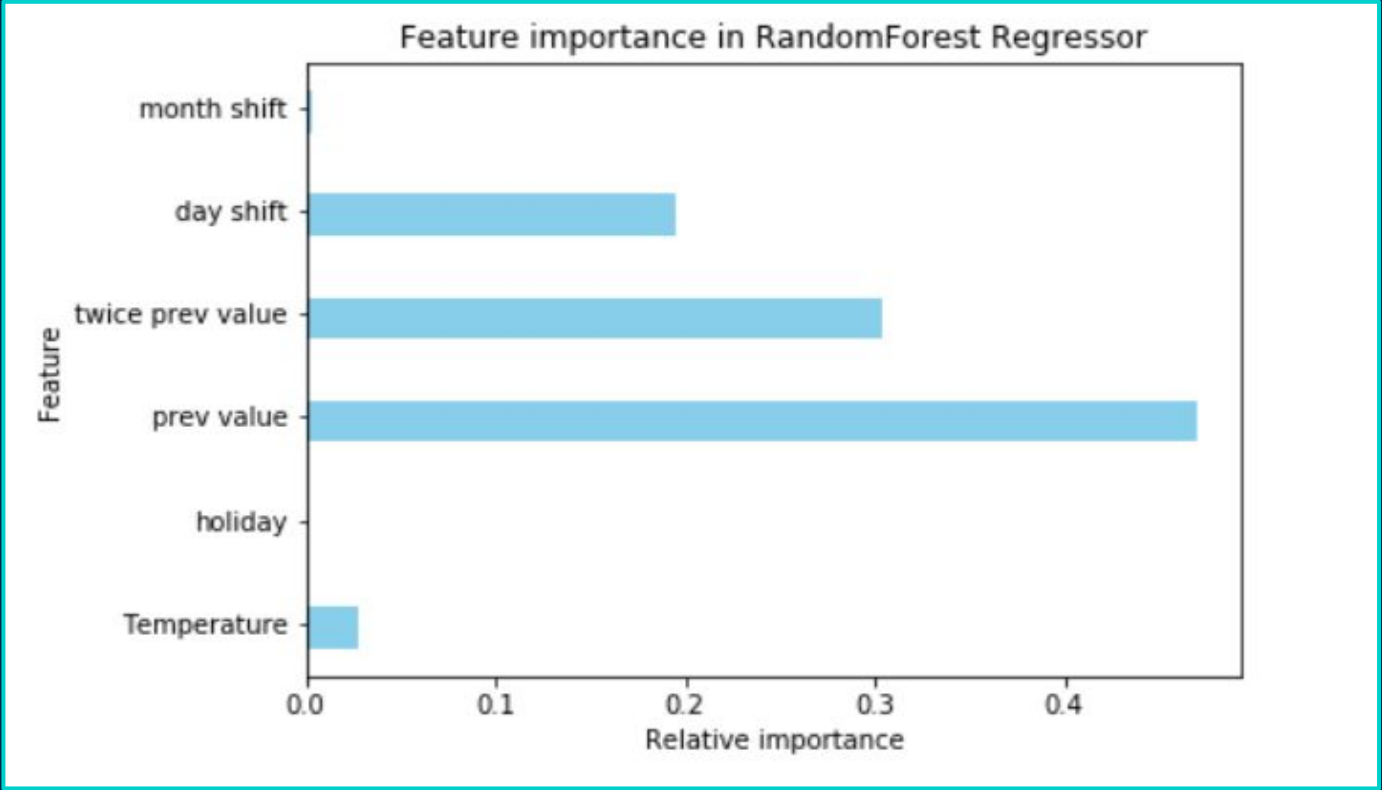
Model Evaluation

	Random Forest	XGBoost	MLP
Train RMSE	15.1757	15.2675	15.2652
Test RMSE	15.1413	15.2236	15.2569
Difference	0.0344	0.0438	0.0083

Time-Value Delta

Min: 0
Max: 524
Mean: 129.25
SD: 48.9

Feature Contribution



ANOMALY DETECTION

Goal

Identify times and attributes that often result in anomalies. Use this information to prevent future excess energy usage.

KNN Anomaly Detection

Isolation Forest Detection

Other Methods for Anomaly Detection

PYOD Anomaly Detection Process

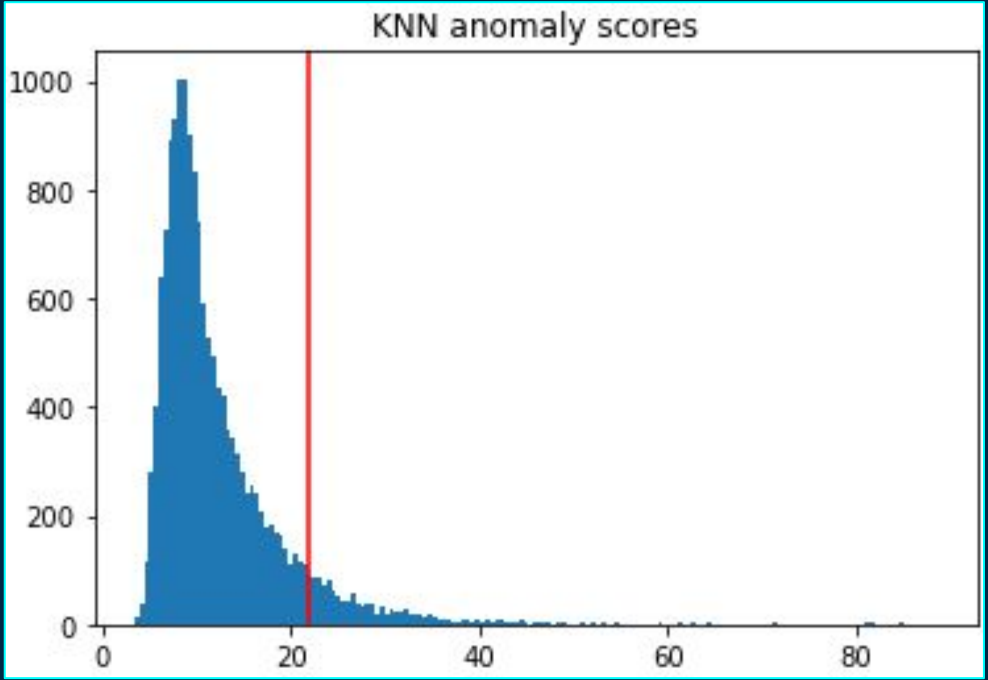
1. Trained **unsupervised model** from PYOD Package on variables in training set
2. Generated **anomaly scores** based on a set criteria for the model
3. Subjective decision, points with **highest scores** are declared outliers
4. Averaged **several models** to reduce overfitting

*Models used included KNN and Isolation Forest

KNN Anomaly Detection – Score Distribution

Averaged anomaly scores from multiple KNN models

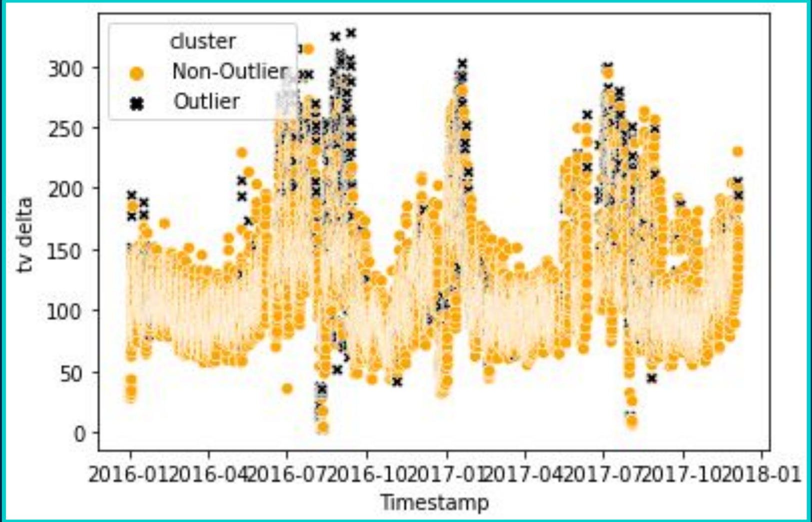
Based on distance between a test point and its Kth nearest neighbor



KNN Anomaly Detection – Data

PYOD models define many high energy values as outliers, without always considering seasonality

KNN model was very dependent on differenced values

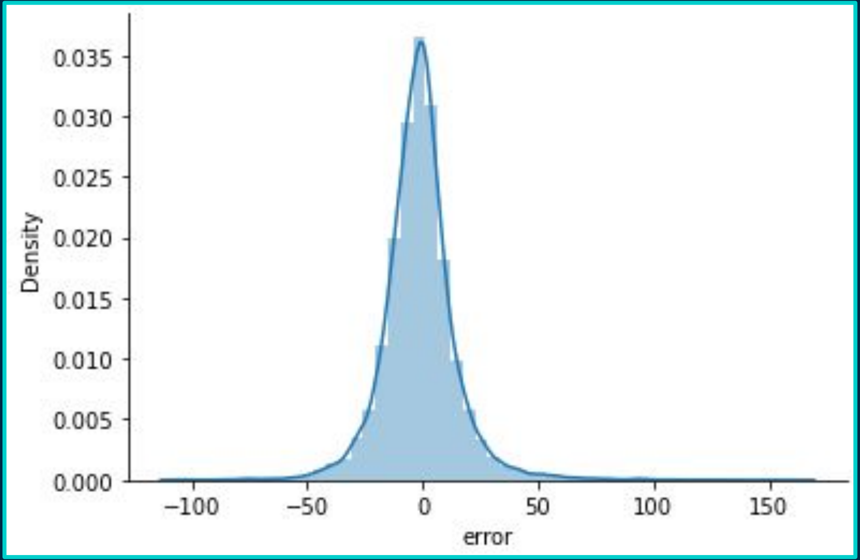


	prev value	twice prev value	tv delta	score
cluster				
Non-Outlier	111.797042	111.985445	112.092058	10.776917
Outlier	161.681620	159.384712	158.393103	30.305915

Anomaly Detection – Random Forest Model

Compared predicted output of our Random Forest model to actual values

Compute moving average of errors over a 2 week period and classify a point as anomalous if the error is **3 SDs** from the mean



Anomaly Detection – Our Model

Date	Actual Values	Predicted	% Difference	Severity (0-3)
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
-03T21:00:0	146	157.51	-7.884%	0
-03T20:00:0	146	133.33	8.678%	0
-03T19:00:0	116	159.057	-37.118%	3
-03T18:00:0	148	161.983	-9.448%	0
-03T17:00:0	157	148.61	5.344%	0
-03T16:00:0	157	132.08	15.873%	1

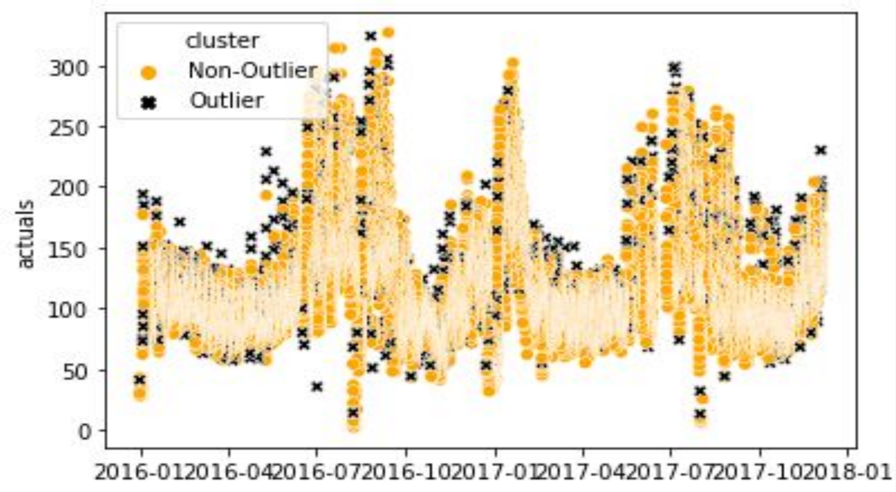
Anomaly Detection - Our Model

Anomaly Detection – Our Model

Difference values still higher on average for outliers, but more reasonable than KNN

~3% of points were outliers with Random Forest

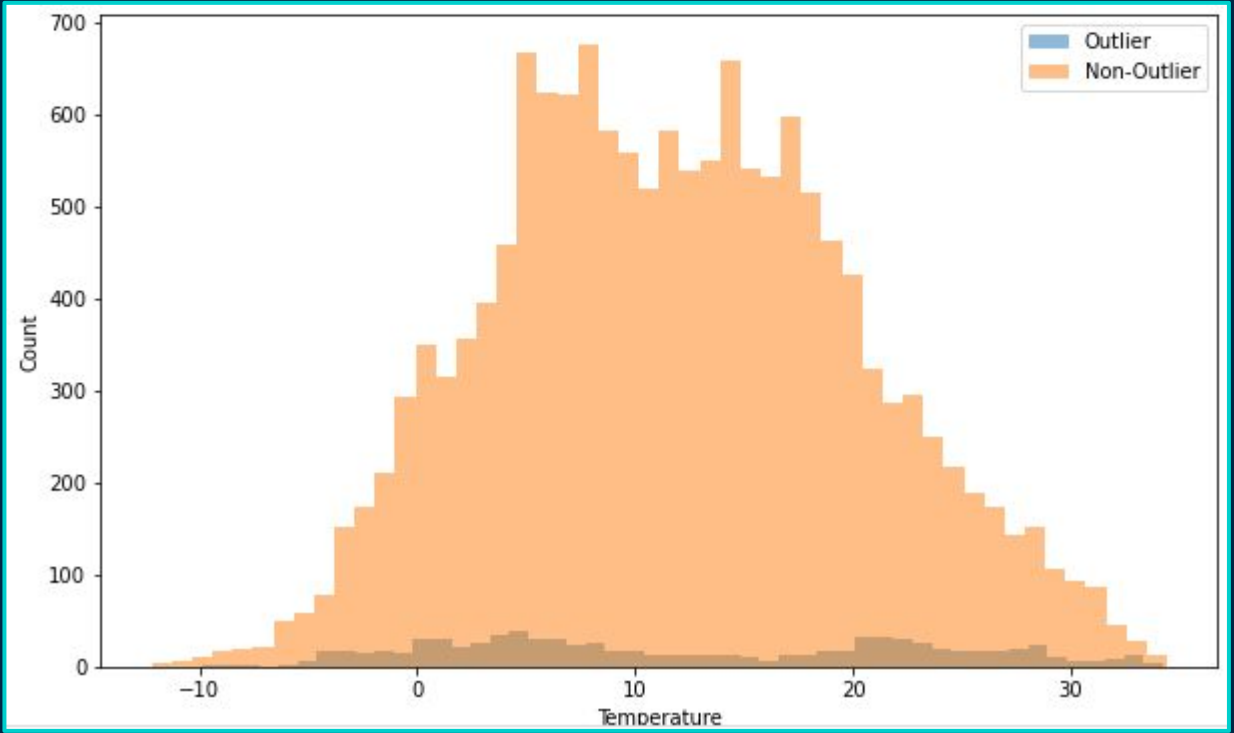
	prev value	twice prev value	actuals
cluster			
Non-Outlier	114.856266	115.066688	114.431438
Outlier	132.190637	128.186796	140.126456



Anomaly Detection – Temperature vs. Outlier

Anomalies seem more frequent at **higher** and **lower** temperatures

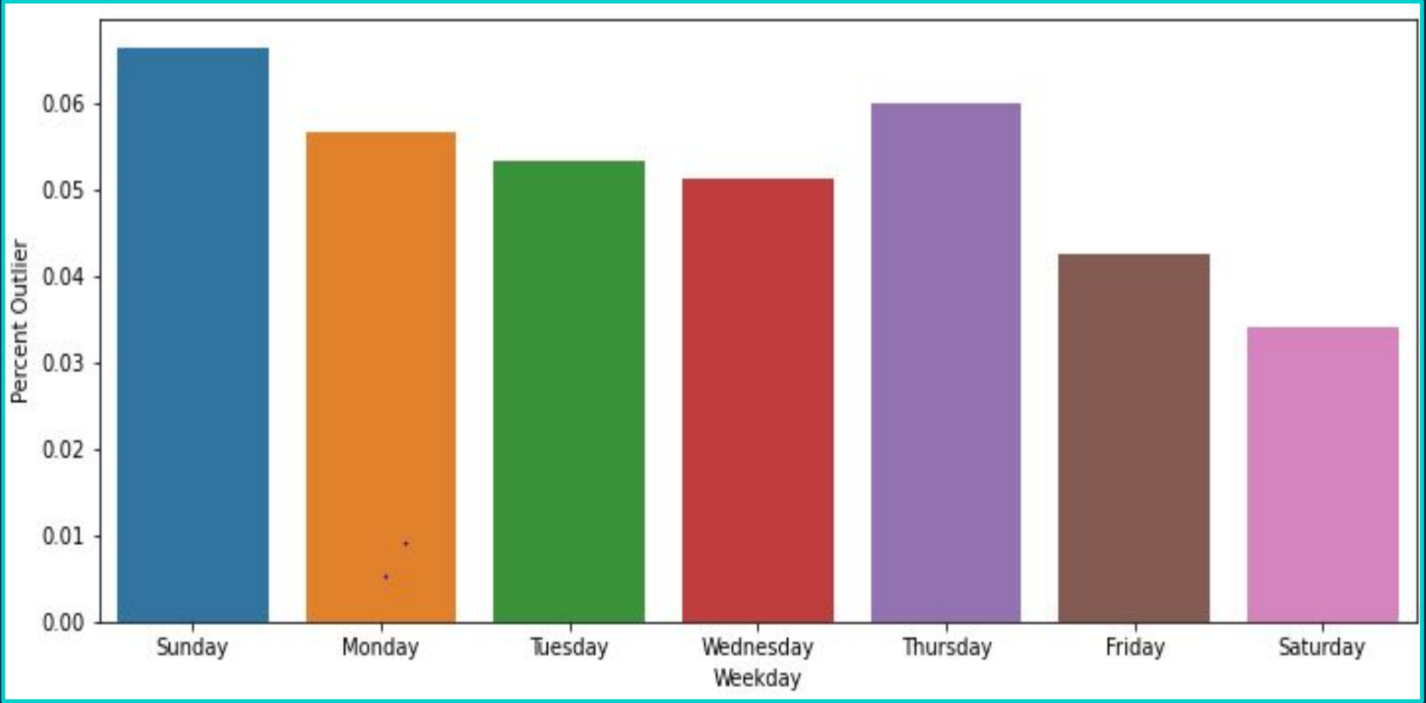
(Mean temperature around 12)



Anomaly Detection – Weekday vs. Outlier

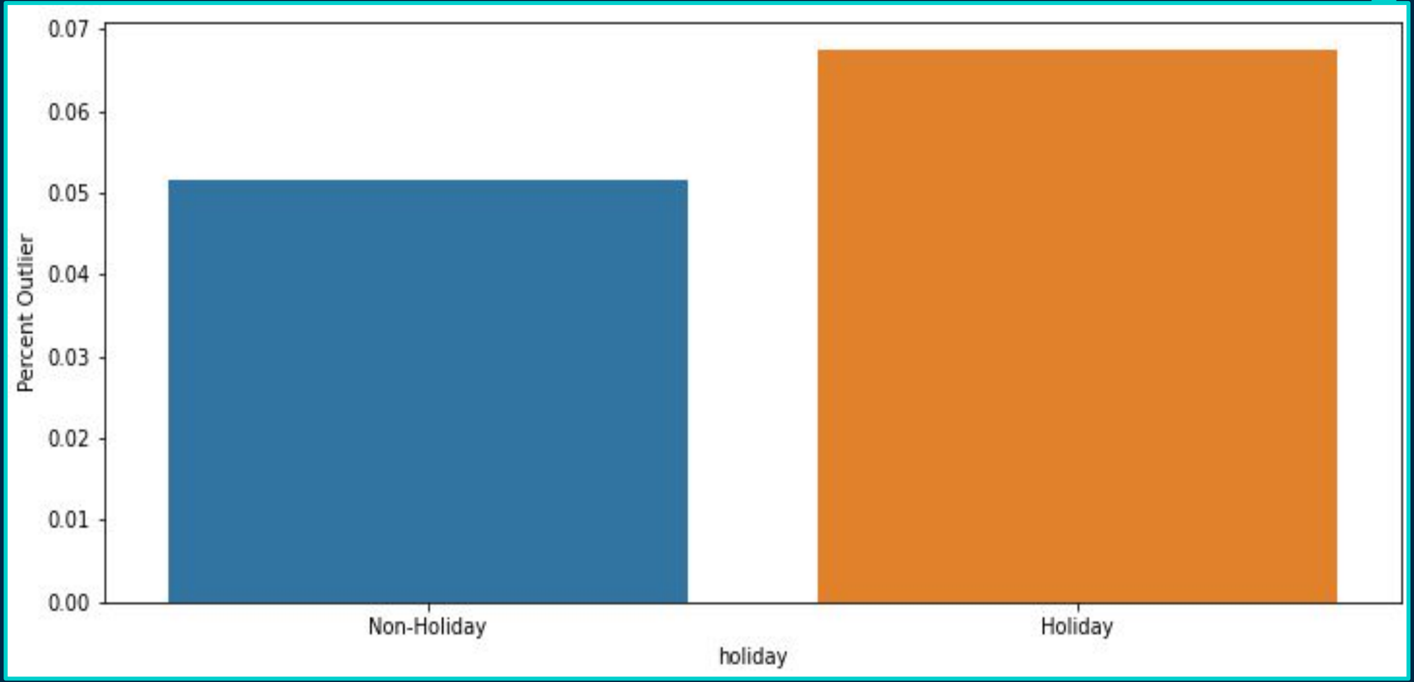
Friday-Sunday are the most different in terms of outlier percentage

Friday and Saturday have lowest percentage, Sunday has highest



Anomaly Detection – Holiday vs. Outlier

There is a higher percentage of anomalous energy usage on Holidays



Takeaways

Understanding **Model Inaccuracies** versus **Data Anomalies**

Dynamic Anomaly Detection

Majority of anomalies occurred during **extreme** temperatures. Firm should adjust for extreme temperatures when managing energy usage during winter and summer months.

Rapid Temperature Changes

More outliers occur when change in temperature changes at an increased rate.

Additional Useful Features

Age and quality of meter, machine and labor hours, intensity of energy use regulations

Application of Model Outcomes

Influence power suppliers’ variable and fixed **pricing** models for corporate energy usage, internal projections for building **costs**, and preventative **maintenance** of meters and other energy equipment

Do you have any questions?

THANK YOU

References

https://www.drivendata.org/competitions/52/anomaly-detection-electricity/
https://github.com/drivendataorg/power-laws-anomalies