Forecasting Future Sales: A Time Series Analysis Utilizing Deep Neural Networks

Montanna Colburn
April 2019
In Collaboration With: Enterprise Training Solutions

Business Problem

Enterprise Training Solutions(ETS) is an e-learning company that sells government training programs for federal, state, local and educational institutions.

Though the company provides e-learning solutions, predicting future sales trends can be tremendously helpful for any business, tangible goods or not. Financial planning, product price points, marketing, hiring needs are just a few benefits from sales forecasting.

Research Questions

Can we predict future sales?

- What time series modeling will best suit our data?
  - How should we forecast, monthly/weekly/daily?
- What attributes can be used to aid in our prediction?
  - If using other attributes, how much of an impact do they have on the model?

Dataset

- The dataset is comprised of 9,977 observations by 23 columns **prior to cleaning*
- Observations from 1997 - 2018
- Attributes **not all 23 columns had inherent value and were removed off the bat*
    - Date (date of sale), Quantity (quantity sold), Type (type of sale), Name State (state where customer is located), Name (customer), Item (the product), Account (supplier), Class (customer type), Rep (sales representative), Sales Price (product price at sale), Amount (total sales amount), Account Type (payment account type), Balance (customers account balance)
- Target variable will be the Quantity of Sales

Pipeline

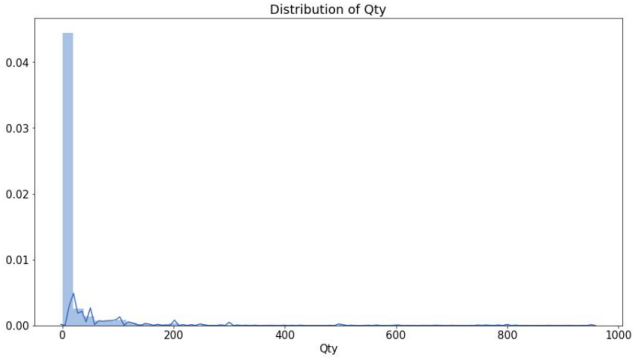
- Imports
- Data Cleaning/ Preprocessing
- Exploratory Data Analysis and Data Visualizations
- Time Series Analysis and Modeling
  - Monthly Aggregated Implementation
  - Daily Aggregated Implementation
- Deep Learning with RNN/ LSTM
  - Monthly Aggregated Implementation
  - Daily Aggregated Implementation
  - Daily Implementation with added Features
- Implementation of Gradient Boosted Decision Tree
  - Feature Analysis
- Final Outcomes
  - Interpretations and Limitations
  - Future Work

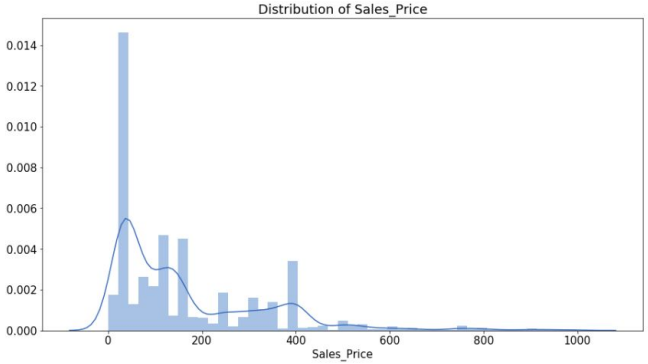
Post Data Cleaning

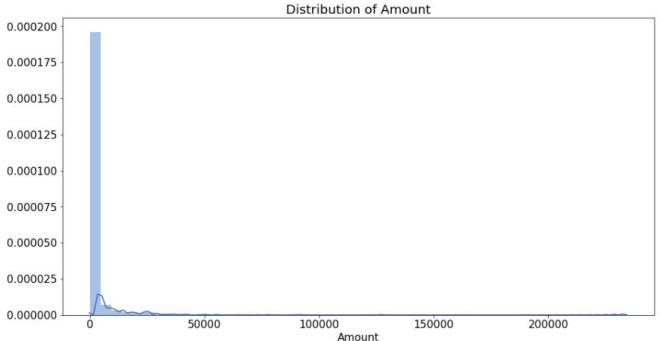
- 8,270 observations available for our Time Series Modeling
- 7,058 observations available for our multi-input Deep Neural Network and Gradient Boosted Model

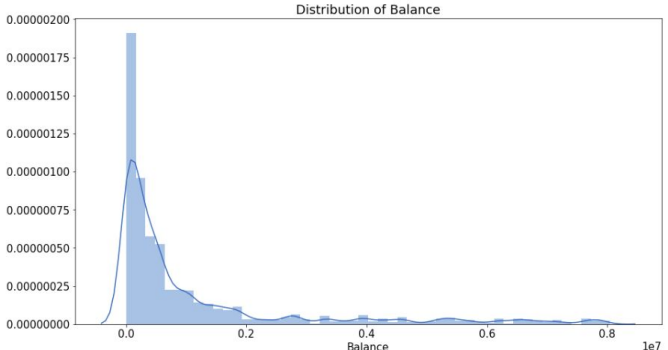
> *Why are these different?* Simply our RNN and GBM will include more features, and in the case of dropped null values, our observations become less.

Data Visualization

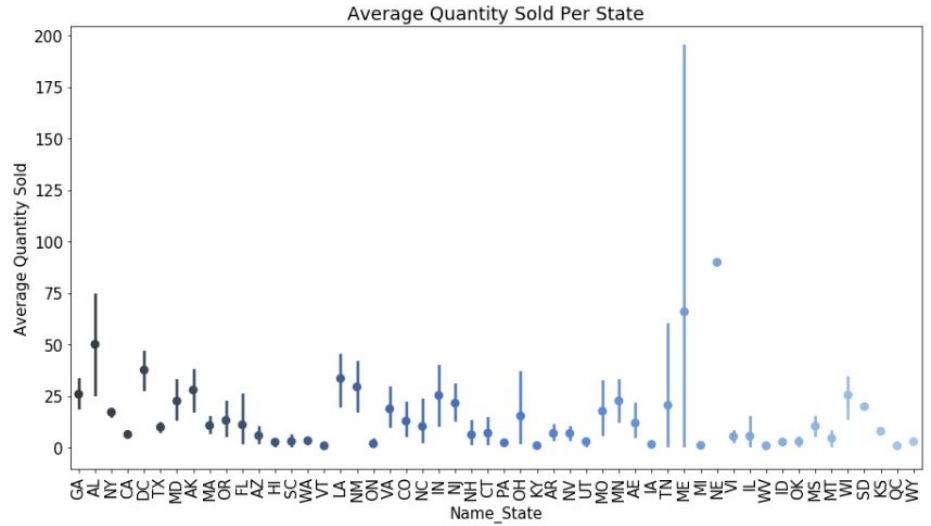


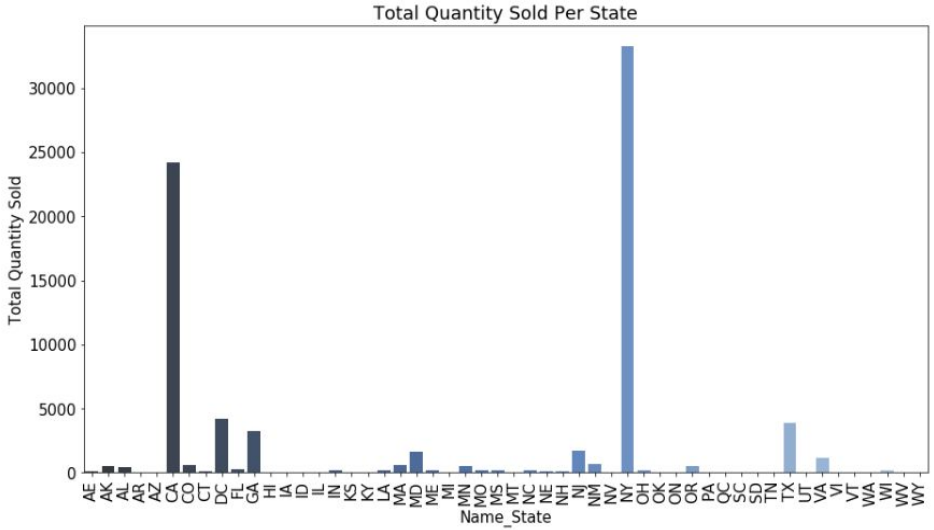




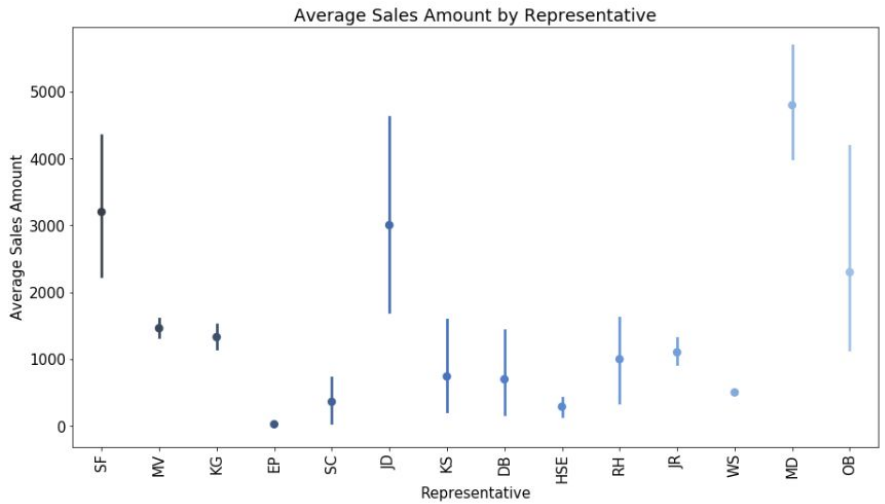


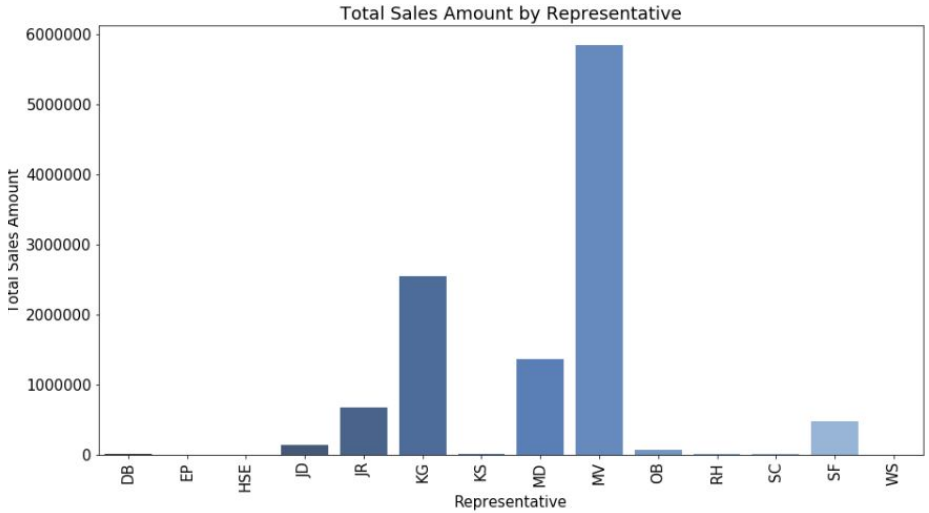
Data Visualization



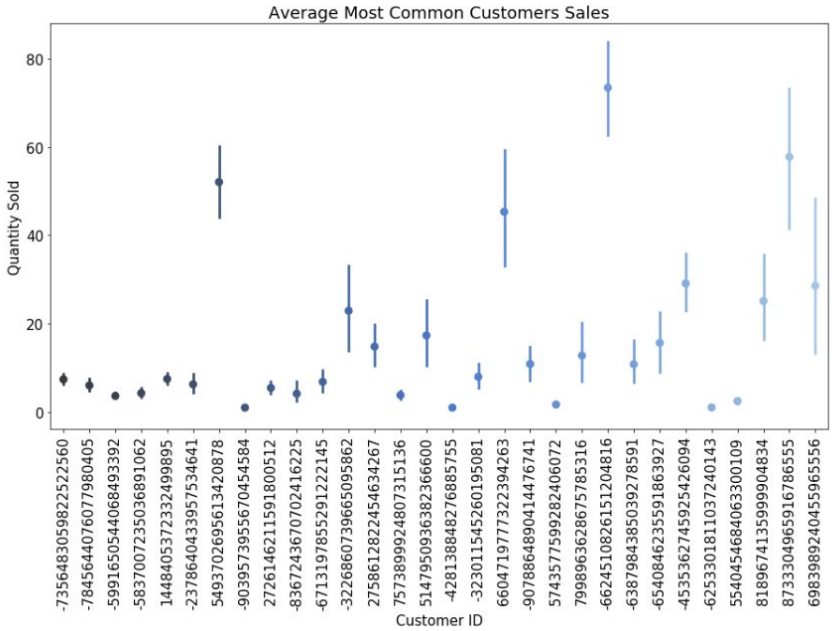


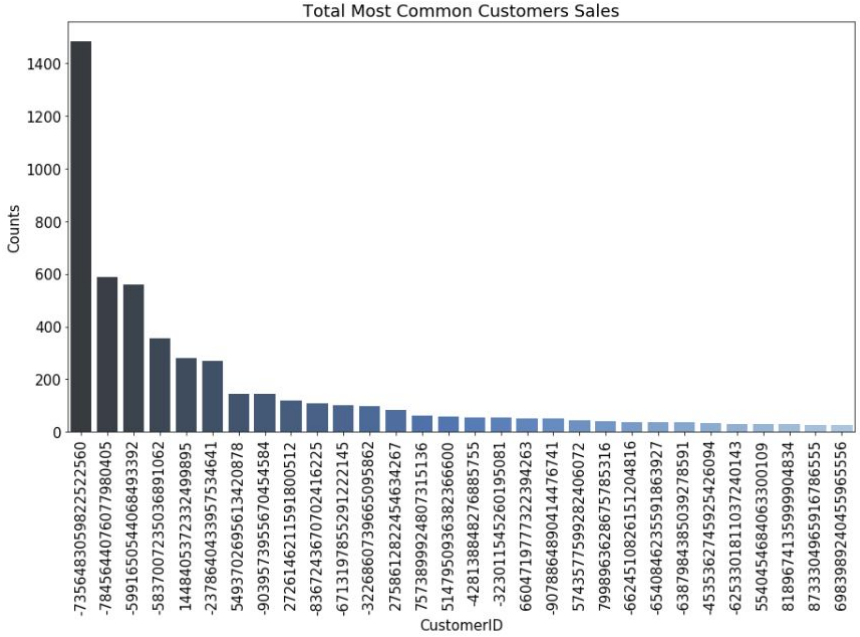
Data Visualization



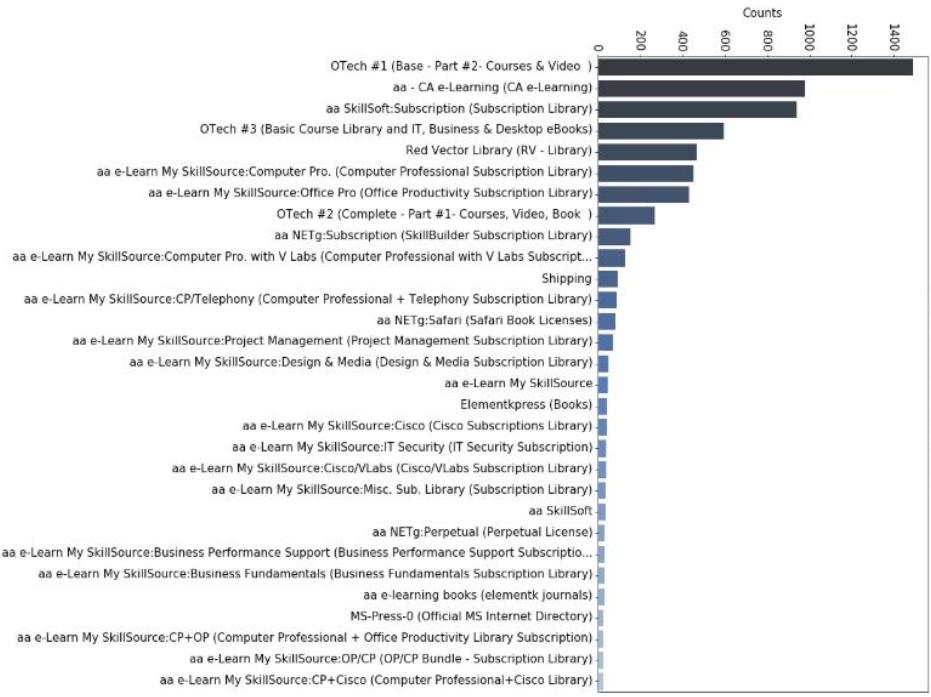


Data Visualization

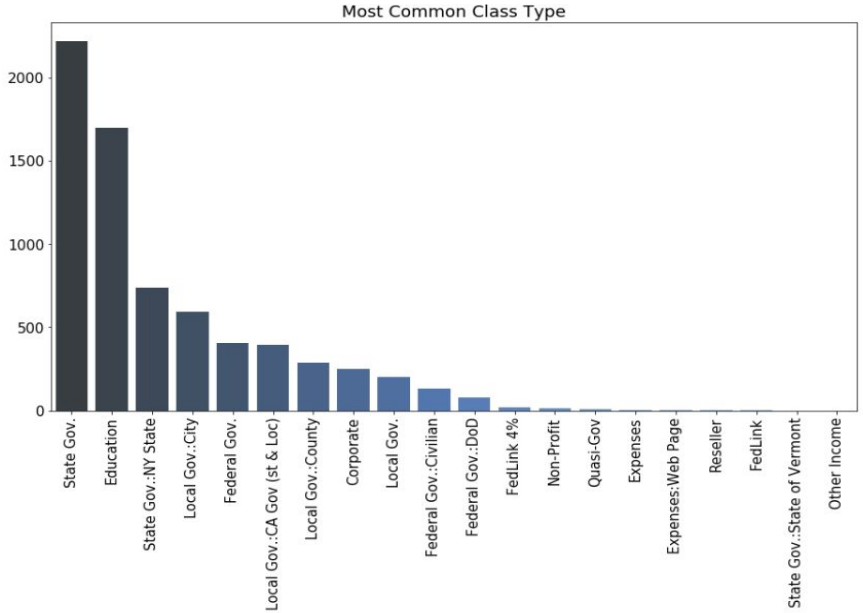


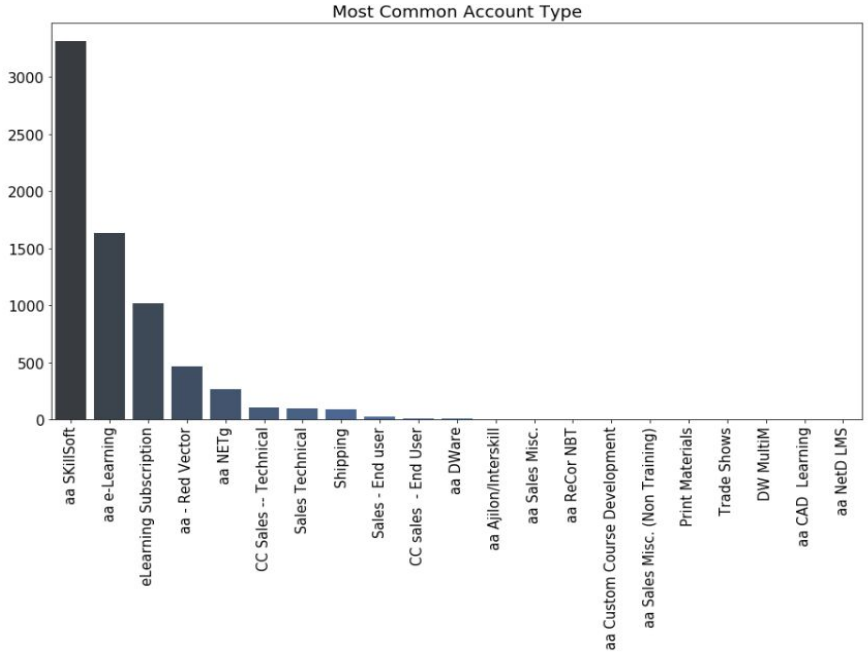



Quantity Sold
0
10
20
30
40
50
OTech #1 (Base - Part #2- Courses & Video)
aa - CA e-Learning (CA e-Learning)
aa SkillSoft:Subscription (Subscription Library)
OTech #3 (Basic Course Library and IT, Business & Desktop eBooks)
Red Vector Library (RV - Library)
aa e-Learn My SkillSource:Computer Pro. (Computer Professional Subscription Library)
aa e-Learn My SkillSource:Office Pro (Office Productivity Subscription Library)
OTech #2 (Complete - Part #1- Courses, Video, Book)
aa NETg:Subscription (SkillBuilder Subscription Library)
aa e-Learn My SkillSource:Computer Pro. with V Labs (Computer Professional with V Labs Subscript...
Shipping
aa e-Learn My SkillSource:CP/Telephony (Computer Professional + Telephony Subscription Library)
aa NETg:Safari (Safari Book Licenses)
aa e-Learn My SkillSource:Project Management (Project Management Subscription Library)
aa e-Learn My SkillSource:Design & Media (Design & Media Subscription Library)
aa e-Learn My SkillSource
Elementkpress (Books)
aa e-Learn My SkillSource:Cisco (Cisco Subscriptions Library)
aa e-Learn My SkillSource:IT Security (IT Security Subscription)
aa e-Learn My SkillSource:Cisco/VLabs (Cisco/VLabs Subscription Library)
aa e-Learn My SkillSource:Misc. Sub. Library (Subscription Library)
aa SkillSoft
aa e-Learn My SkillSource:Business Performance Support (Business Performance Support Subscriptio...
aa NETg:Perpetual (Perpetual License)
aa e-Learn My SkillSource:Business Fundamentals (Business Fundamentals Subscription Library)
aa e-learning books (elementk journals)
MS-Press-0 (Official MS Internet Directory)
aa e-Learn My SkillSource:CP+OP (Computer Professional + Office Productivity Library Subscription)
aa e-Learn My SkillSource:OP/CP (OP/CP Bundle - Subscription Library)
aa e-Learn My SkillSource:CP+Cisco (Computer Professional+Cisco Library)
Average Most Common Product Sales


Counts
0
200
400
600
800
1000
1200
1400
OTech #1 (Base - Part #2- Courses & Video)
aa - CA e-Learning (CA e-Learning)
aa SkillSoft:Subscription (Subscription Library)
OTech #3 (Basic Course Library and IT, Business & Desktop eBooks)
Red Vector Library (RV - Library)
aa e-Learn My SkillSource:Computer Pro. (Computer Professional Subscription Library)
aa e-Learn My SkillSource:Office Pro (Office Productivity Subscription Library)
OTech #2 (Complete - Part #1- Courses, Video, Book)
aa NETg:Subscription (SkillBuilder Subscription Library)
aa e-Learn My SkillSource:Computer Pro. with V Labs (Computer Professional with V Labs Subscript...
Shipping
aa e-Learn My SkillSource:CP/Telephony (Computer Professional + Telephony Subscription Library)
aa NETg:Safari (Safari Book Licenses)
aa e-Learn My SkillSource:Project Management (Project Management Subscription Library)
aa e-Learn My SkillSource:Design & Media (Design & Media Subscription Library)
aa e-Learn My SkillSource
Elementkpress (Books)
aa e-Learn My SkillSource:Cisco (Cisco Subscriptions Library)
aa e-Learn My SkillSource:IT Security (IT Security Subscription)
aa e-Learn My SkillSource:Cisco/VLabs (Cisco/VLabs Subscription Library)
aa e-Learn My SkillSource:Misc. Sub. Library (Subscription Library)
aa SkillSoft
aa NETg:Perpetual (Perpetual License)
aa e-Learn My SkillSource:Business Performance Support (Business Performance Support Subscriptio...
aa e-Learn My SkillSource:Business Fundamentals (Business Fundamentals Subscription Library)
aa e-learning books (elementk journals)
MS-Press-0 (Official MS Internet Directory)
aa e-Learn My SkillSource:CP+OP (Computer Professional + Office Productivity Library Subscription)
aa e-Learn My SkillSource:OP/CP (OP/CP Bundle - Subscription Library)
aa e-Learn My SkillSource:CP+Cisco (Computer Professional+Cisco Library)
Total Most Common Product Sales

Data Visualization

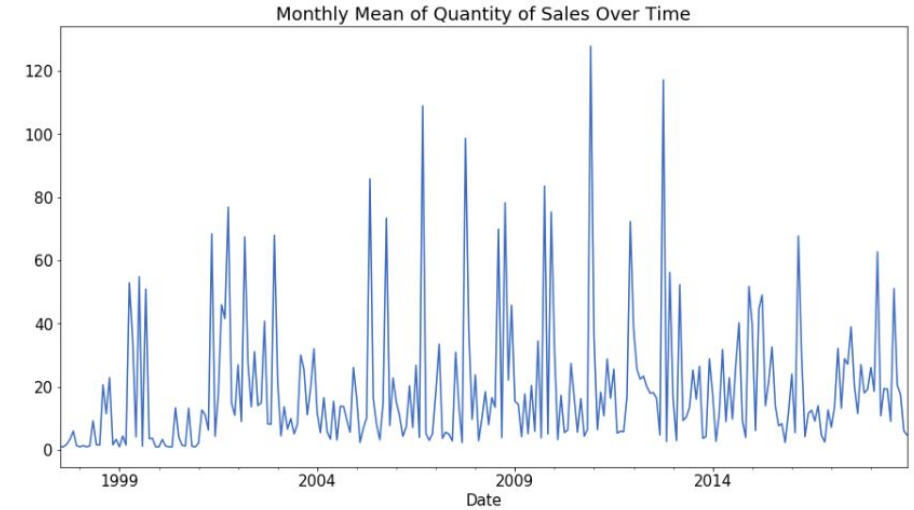


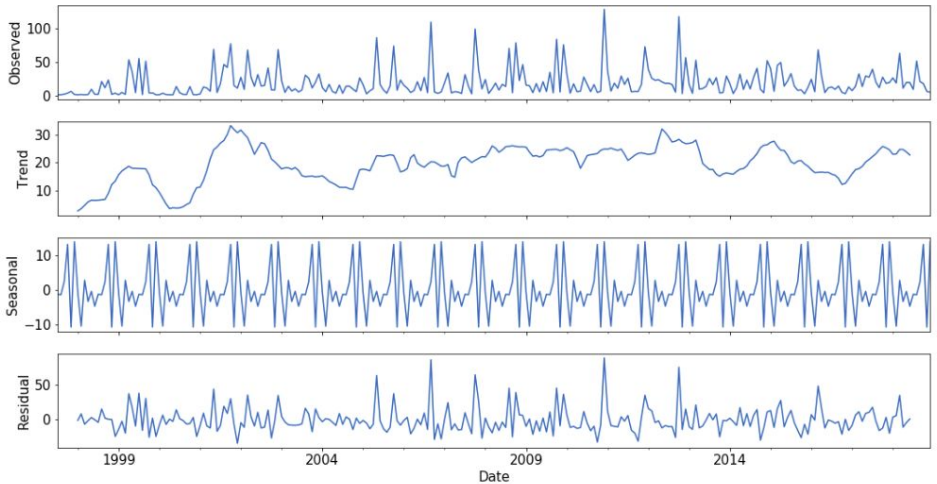


Time Series ARIMA Modeling

- Monthly Aggregated Implementation
- Daily Aggregated Implementation

Monthly ARIMA Prediction



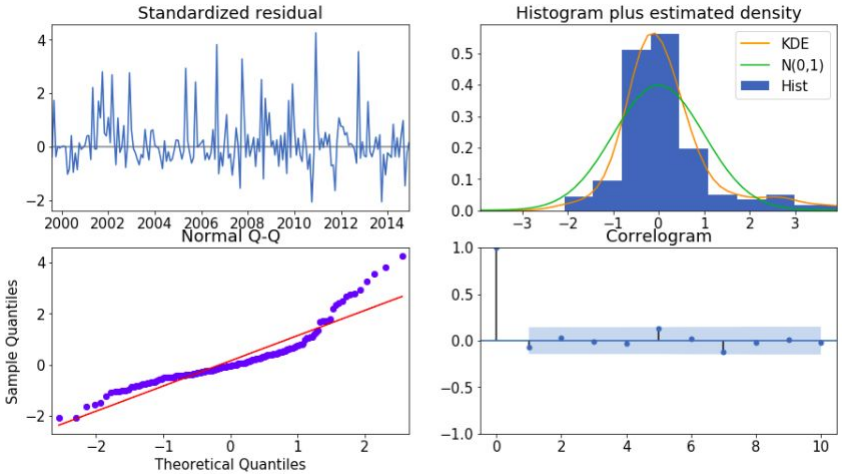


Monthly ARIMA Prediction

Second Best AIC Value: 2104.5863324367856.

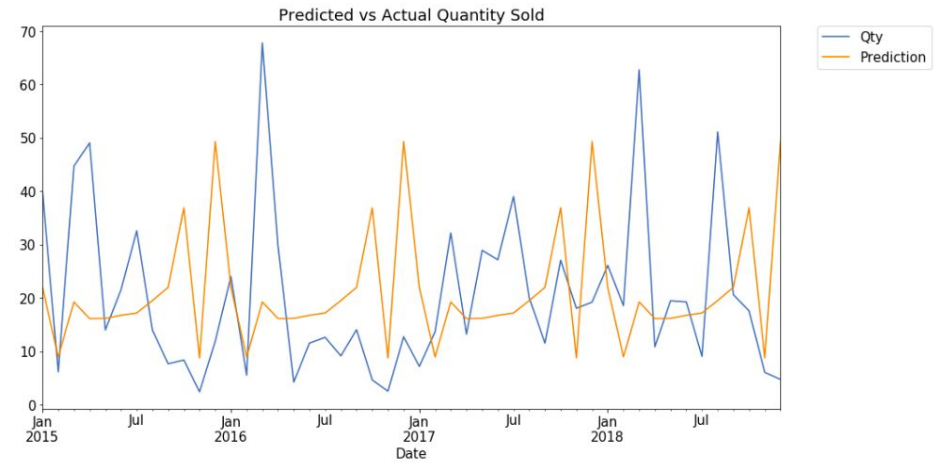
```
mod2 = sm.tsa.statespace.SARIMAX(train,
                                 order=(0, 0, 0),
                                 seasonal_order=(0, 1, 1, 12),
                                 enforce_stationarity=False,
                                 enforce_invertibility=False)
results2 = mod2.fit()
print(results2.summary().tables[1])
```

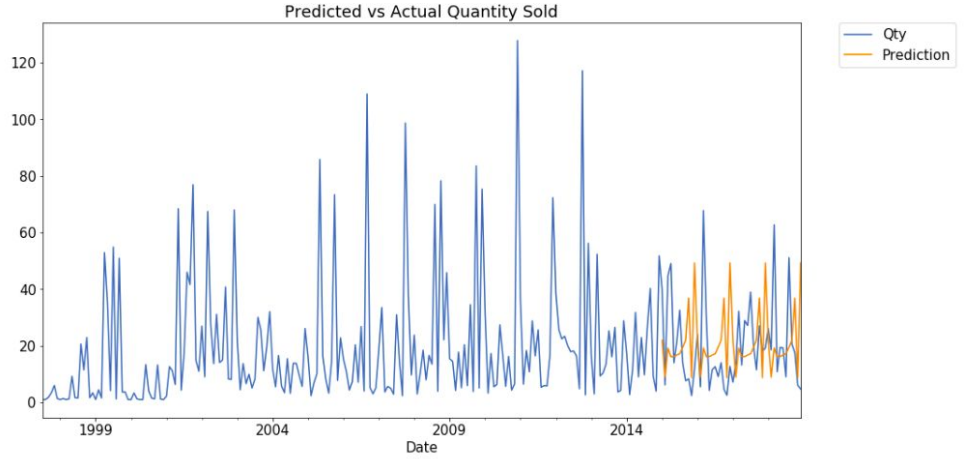
```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.S.L12      -0.8299      0.040    -20.853      0.000      -0.908      -0.752
sigma2       511.2685     30.522     16.751      0.000     451.447     571.090
==============================================================================
```



Monthly ARIMA Prediction

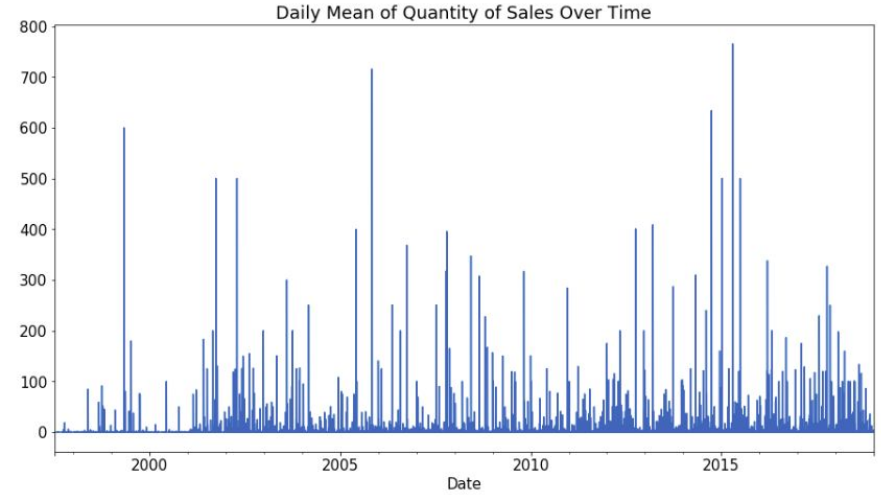
RMSE: 3.7986807331779495

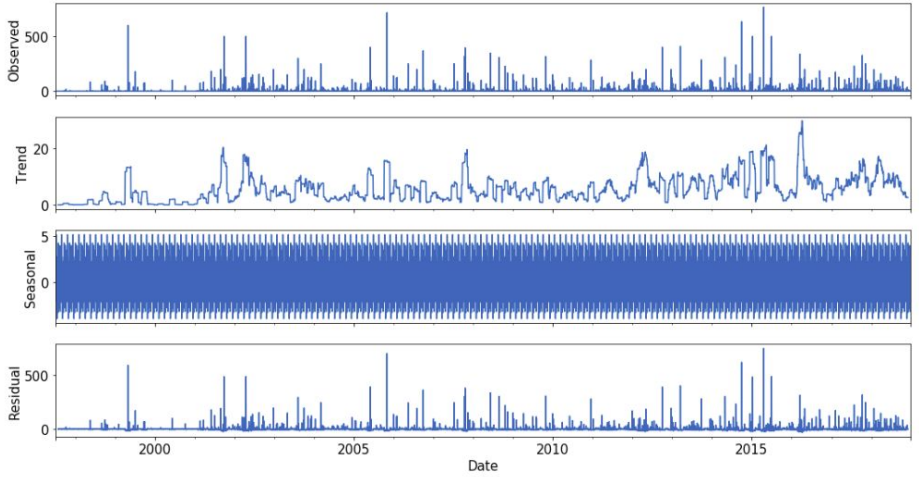




Daily ARIMA Prediction

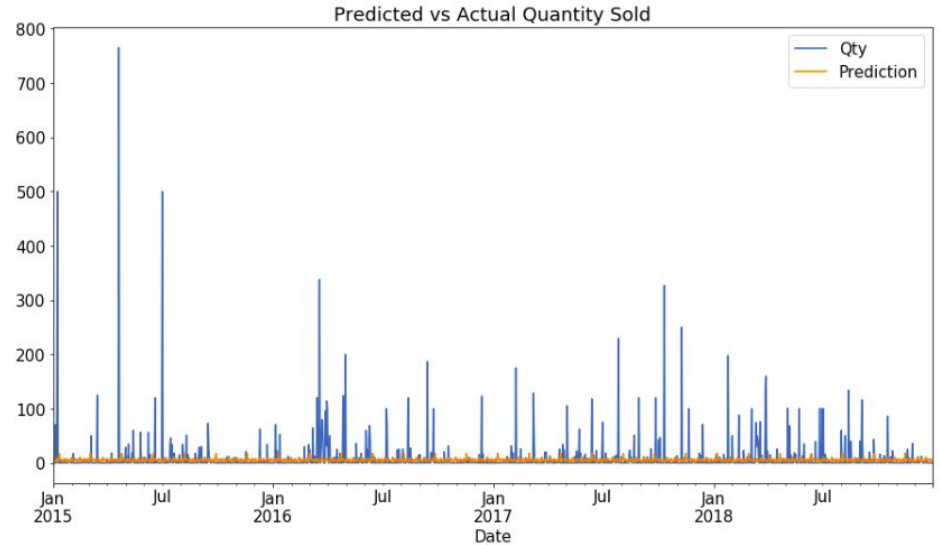
- Downside of SARIMAX modeling is its incapability of memory storage to execute a 365 seasonality prediction. Instead, I used 52 for weekly seasonality, yet still ran into computation storage issues.

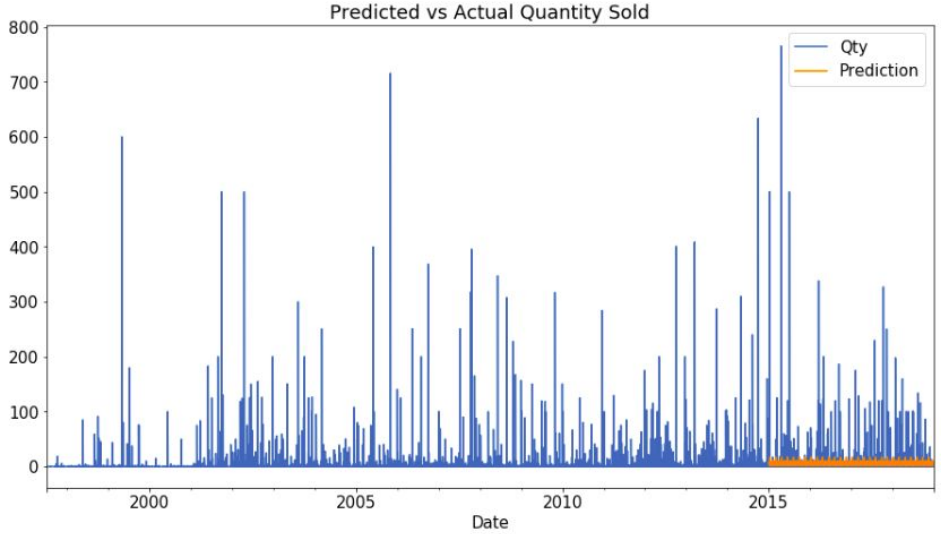




Daily ARIMA Prediction

RMSE: 3.39238159410763

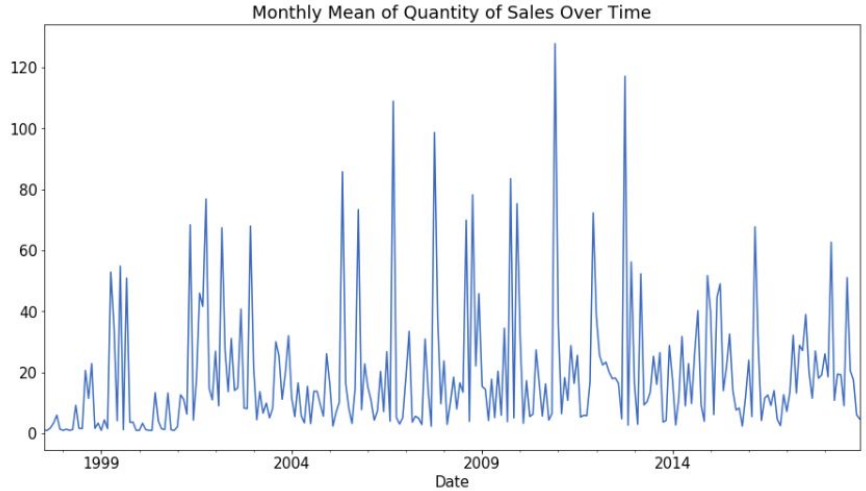




LSTM Modeling

- Monthly Aggregated Implementation
- Daily Aggregated Implementation
- Daily Implementation with added Features

Monthly LSTM Prediction

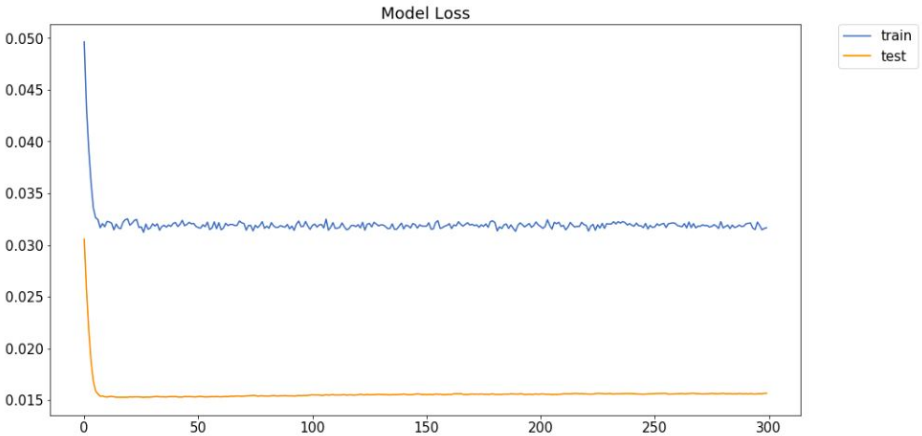


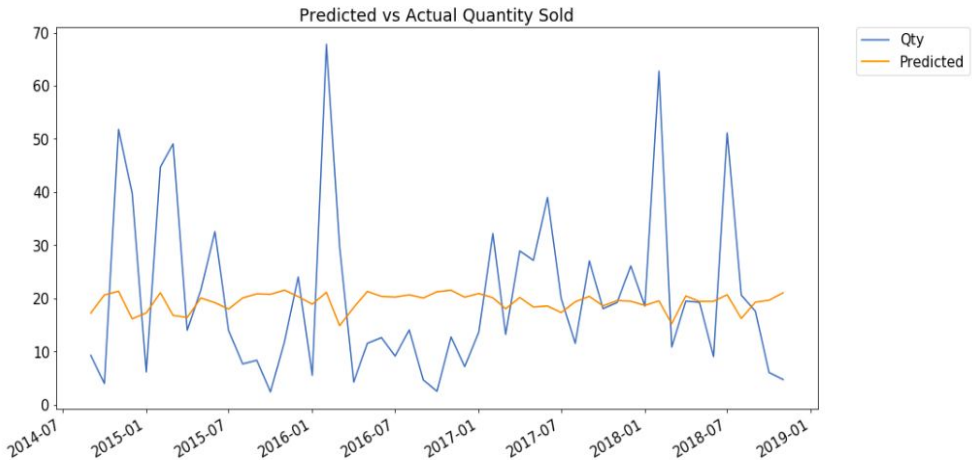

```
# The LSTM architecture
model = Sequential()
# First LSTM layer with Dropout regularization
model.add(LSTM(units=50, return_sequences=True, input_shape=(trainX.shape[1], trainX.shape[2])))
model.add(Dropout(0.2))
# Second LSTM layer
model.add(LSTM(units=50, return_sequences=True))
model.add(Dropout(0.2))
# Third LSTM layer
model.add(LSTM(units=50, return_sequences=True))
model.add(Dropout(0.2))
# Fourth LSTM layer
model.add(LSTM(units=50))
model.add(Dropout(0.2))
# The output layer
model.add(Dense(units=1))

# Compiling the RNN
model.compile(optimizer='rmsprop',loss='mean_squared_error')
# Fitting to the training set
history = model.fit(trainX, trainY,epochs=300,batch_size=100,validation_data=(testX, testY),verbose=0,shuffle=False)
```

Monthly LSTM Prediction

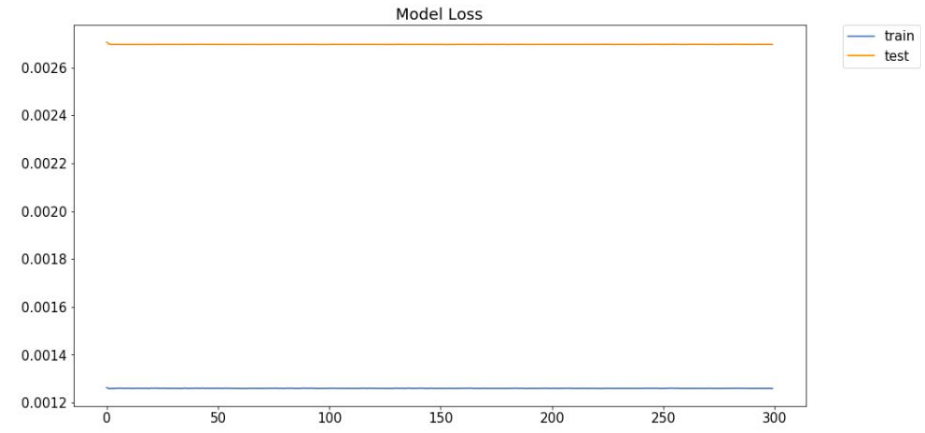
RMSE: 15.835

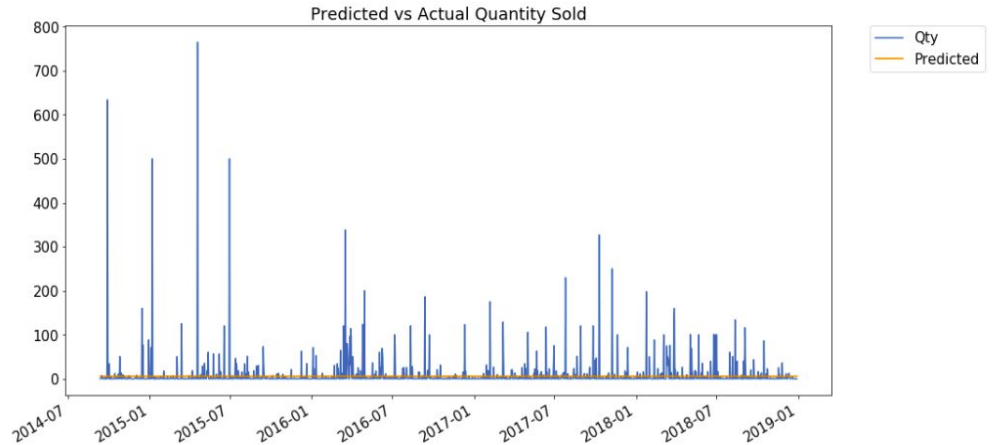




Daily LSTM Prediction

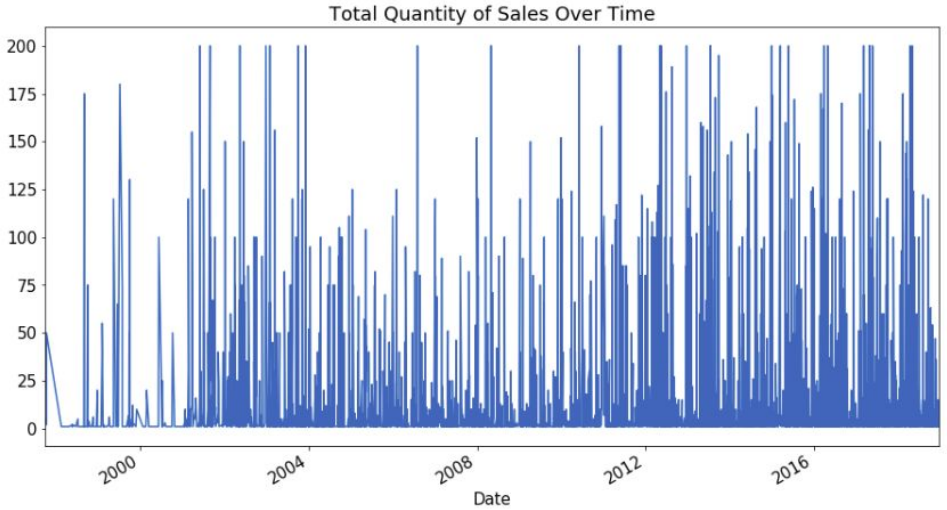
RMSE: 39.720





Daily LSTM Prediction: Added Features

- In our next series of predictions we incorporate known features, and thus, we're asking a new question; *can we predict the quantity sold in a given transaction?*

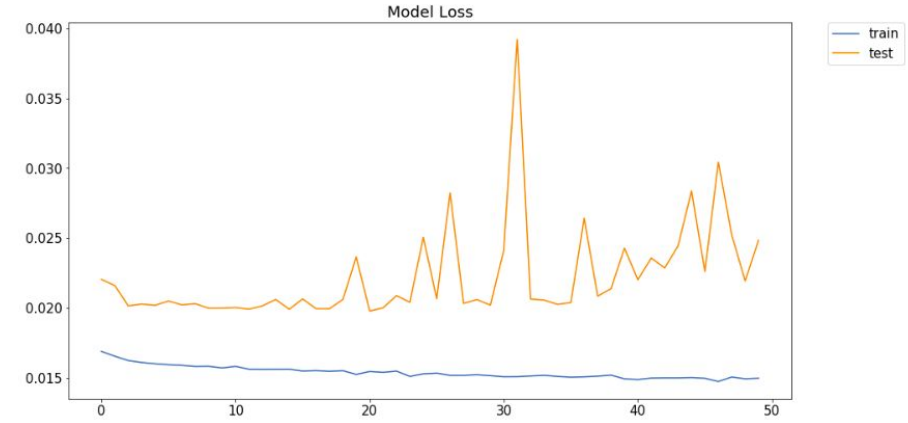


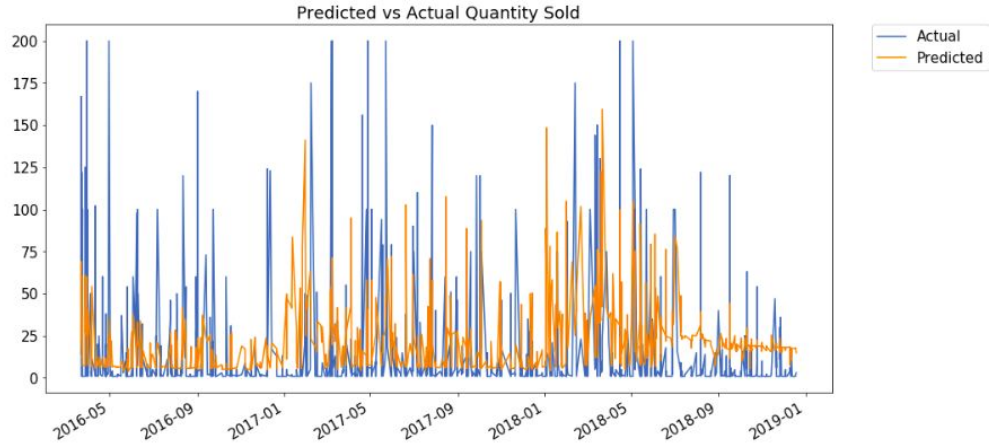
Daily LSTM Prediction: Added Features

- Label encode categorical features
  - Type: 2
  - Name State: 52
  - Item: 183
  - Account: 21
  - Class: 20
  - Rep: 14
  - Account Type: 3
- Lag variable for previous days quantity
- Day/Month/Year variable

Daily LSTM Prediction: Added Features

RMSE: 31.356





XGBoost Modeling

- Daily Implementation
- Feature Analysis

Feature Engineering

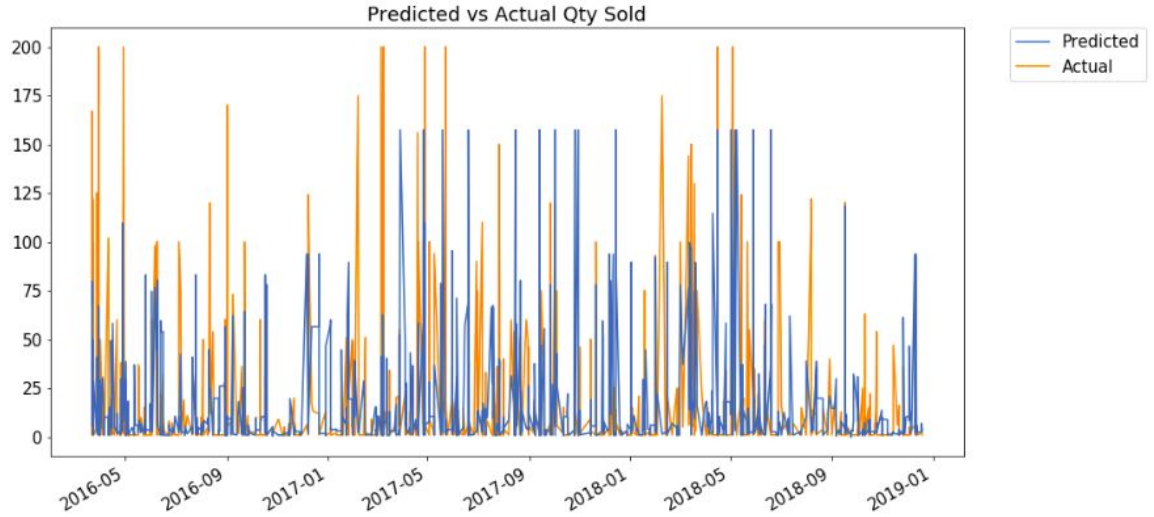
Same as LSTM daily prediction with added features

- Label encode categorical features
  - Type: 2
  - Name State: 52
  - Item: 183
  - Account: 21
  - Class: 20
  - Rep: 14
  - Account Type: 3
- Lag variable for previous days quantity
- Day/Month/Year variable

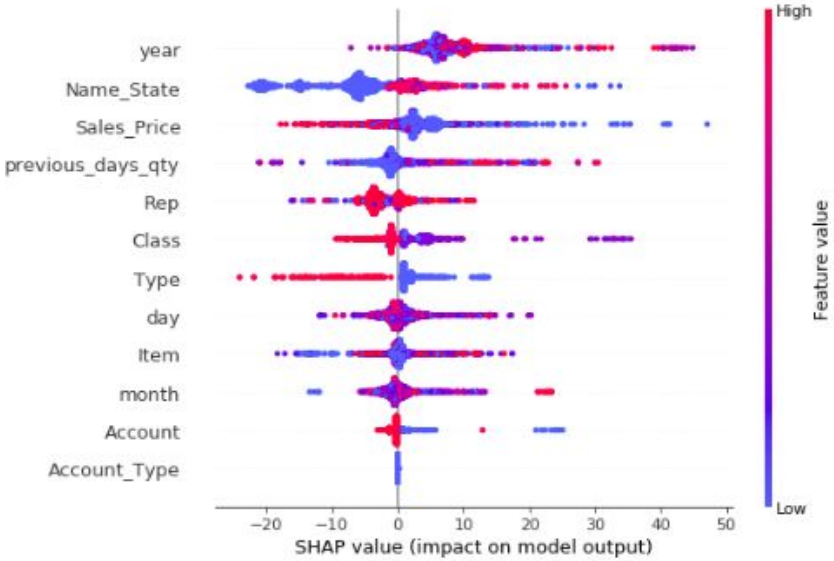
Daily Prediction

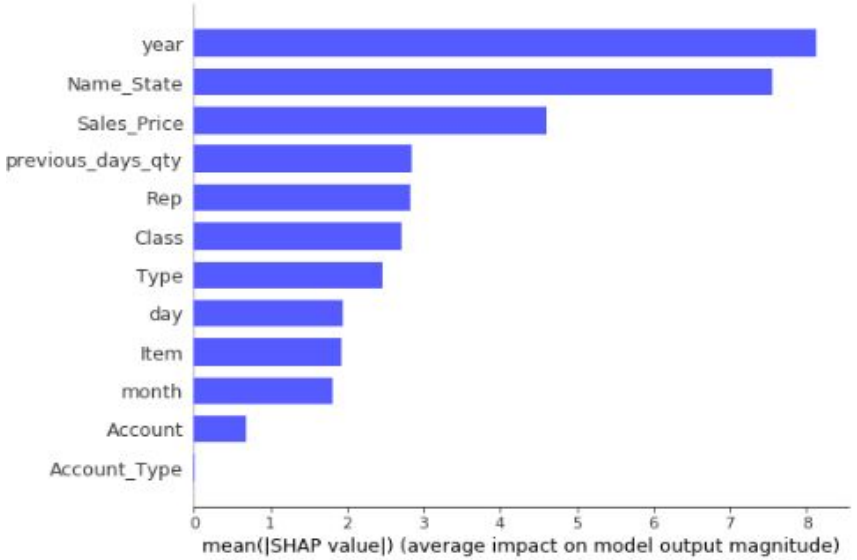
Tuned Parameters: 'learning_rate': 0.1, 'max_depth': 20, 'n_estimators': 20

RMSE: 31.719



Feature Importance





Final Outcomes

Monthly SARIMAX Prediction: Test RMSE: **3.798**

Daily SARIMAX Prediction: Test RMSE: **3.392**

Monthly LSTM Prediction: Test RMSE: **15.835**

Daily LSTM Prediction: RMSE: **39.720**

Daily LSTM Prediction with Added Features: Test RMSE: **31.356**

Implementation of Gradient Boosted Decision Tree: Test RMSE: **31.719**

Further Exploration

- Incorporate more lagged features for the RNN and GBM; i.e previous 3 or 7 day quantity rolling average or other variables shifted values
- Research walk forward validation for LSTMs to see if it'd perform better
- Test out further architectures for LSTM model

Thanks!

Special thanks to ETS for their willingness to give a random samaritan their data, Max Sop for being a great mentor and resource, and the Thinkful community.