1 Overview

The goal is to be able to predict the future direction (and magnitude) of change in the difference between yields of risky and risk-free securities, known as the credit spread. Our focus here will be on predicting the spread of investment-grade US corporate bonds over US treasury bonds, using past information about the credit spread along with assorted exogenous data to construct models for how the credit spread changes over time. We then extrapolate from these models to make claims about the credit spread's future movement. The two main learning frameworks we used were Support Vector Machines and Long Short-Term Memory Neural Networks. We also developed an ARIMA model of the data to use as a benchmark for our learning models.

2 Description of the Data

2.1 Credit Spread Data

All of the predictive models train on the difference between corporate bond and treasury bond yields for various different maturities. The bond yield data was all obtained from the St. Louis Federal Reserve's FRED Economic Database. The corporate bond data was taken from the ICE Bank of America Merrill Lynch Corporate Effective Yield for 1 to 3 year, 3 to 5 year, 5 to 7 year, 7 to 10 year, and 10 to 15 year maturities. The treasury bond data was taken from the 1 year, 2 year, 3 year, 5 year, 7 year, 10 year, and 30 year Treasury Constant Maturity Rates, in addition to the 1 month, 3 month, and 6 month Treasury Constant Maturity Rates.

Our first step towards computing the spread was normalizing the data arranging it into four groups based on bond maturity: 1 to 3 years, 3 to 5 years, 5 to 7 years, and 7 to 10 years. Since the corporate and treasury data for the given maturity groups have different granularity, the spread could not be computed as a pointwise difference. To compute the difference in the yields for each maturity group, we used numerical integration to find the area between the two yield curves. The result of these transformations is four sequences of time-series data of length 5509, one for each maturity group. The data begins on the 31$^{\text{st}}$ of December, 1996, ends on the 17$^{\text{th}}$ of January, 2019, and every time-step in the data is one business day. These time-series data were used to train the models used to predict the spread.

2.2 Exogenous Data

Some of our learning models (namely, the Neural Network) are capable of easily incorporating potentially irrelevant data into the model and autonomously weighing its influence so that the credit spread is optimally predicted. Therefore, we also make use of Dow Jones, S&P 500, NASDAQ, and VIX data from 1997 and inflation rate data from 2000 to the present.

3 Training and Testing the Learning Models

3.1 ARIMA

Our baseline model is the Autoregressive Integrated Moving Average. The ARIMA model was trained exclusively on the credit spread data, incorporating no exogenous information. The model is first fit to the first 4509 *training* elements of the credit spread time-series. Afterwards, the model was iterated on the remaining 1000 *validation* points. The iteration consists of predicting one new

value at a time, adding it to the training set, and refitting the model on the newly expanded training set. This is a fairly slow model to train and predict with, but it gives fairly good results as summarized in Table 1. The following figures plot the *validation* data points against the predicted ARIMA sequence during the iteration.

Table 1: ARIMA Root Mean Squared Error

Series	RMSE
1 to 3 year spread	0.013
3 to 5 year spread	0.014
5 to 7 year spread	0.014
7 to 10 year spread	0.015

Figure 1: ARIMA 1 to 3 year spread

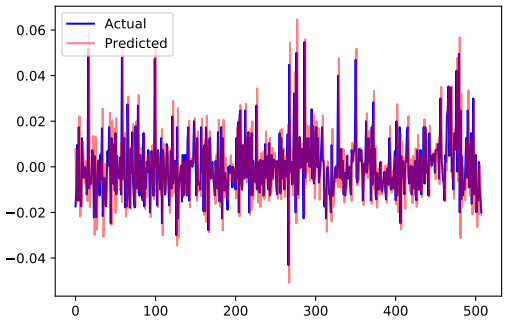


Figure 2: ARIMA 3 to 5 year spread

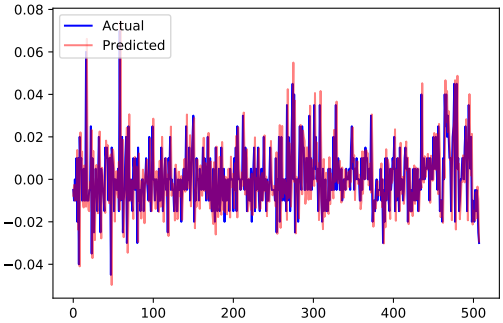


Figure 3: ARIMA 5 to 7 year spread

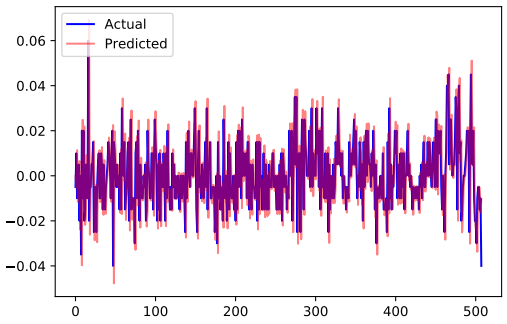
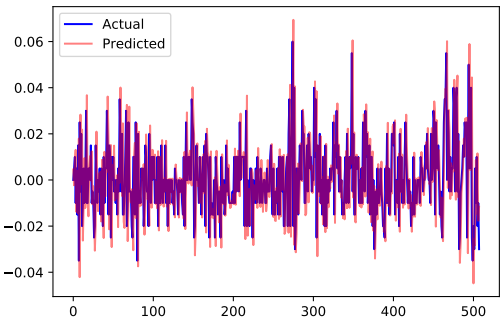


Figure 4: ARIMA 7 to 10 year spread



3.2 Support Vector Machines

Talk about the SVM.

Figure 5: ARIMA 1 to 3 year spread

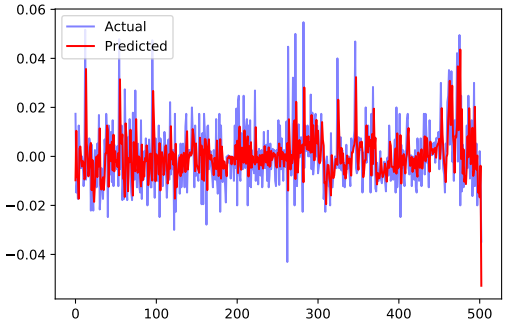


Figure 6: ARIMA 3 to 5 year spread

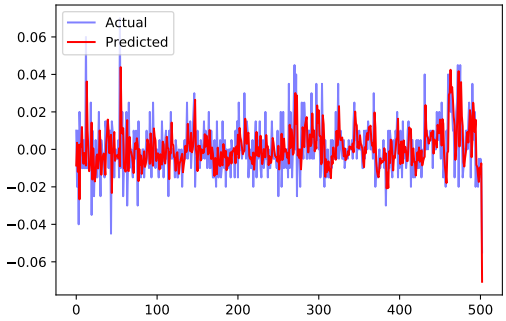


Figure 7: ARIMA 5 to 7 year spread

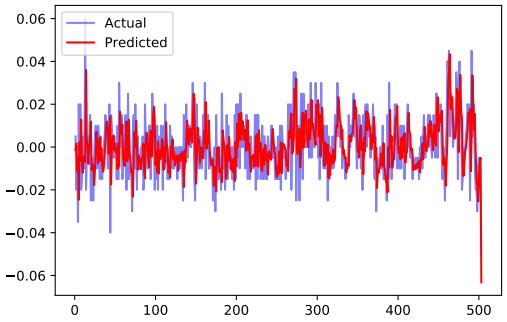
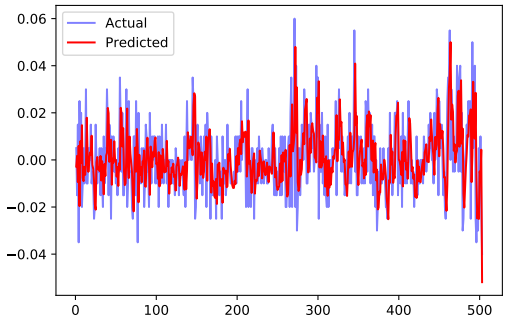


Figure 8: ARIMA 7 to 10 year spread



As we can see, the SVM model is predicting a high-magnitude downward trend at the end of the *validation* set for all of the maturity groups.

3.3 LSTM Neural Networks

Talk about the LSTM NN.

4 Summary of Results

5 Possible Future Directions for Research