Prediction Engine for Quarterly Earnings Using Alternative Data

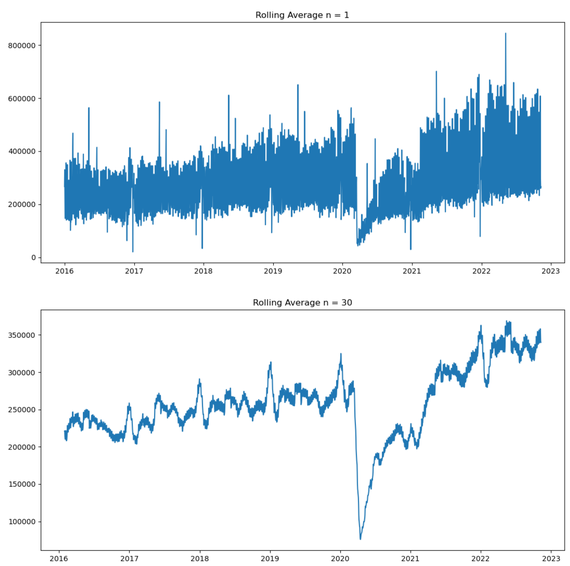
Research: Theo Bazille, Yunqing Cao, Charles Kanter

Overview

Machine learning models to predict company quarterly earning metrics using a credit card transaction dataset of different sampling frequency. Our focus is to detect anomalies in our datasets and track inflection points. We explore correlations between the two datasets and how to best deal with varying frequencies without loss of information (e.g. upsampling). We explore ML/DL models for best prediction and ultimately find that Ridge Regression yields the best prediction accuracy.

Data Preprocessing

- **Data Set 1:** Sequential data consisting of 116,659 rows of daily credit card transaction totals ($USD)
- **Data Set 2:** Sequential data consisting of 781 rows of quarterly sales reports ($USD), same companies from Data Set 1
- **Time Period:** 1/1/2016 to 7/4/2022
- **Unique Companies:** 37
- **Moving Average:** 30-day window (see below)



Visualization

Correlation varied significantly for different companies, as seen below:

Top to bottom: Strong (0.68), weak (0.45), uninterpretable (-0.15)

Upsampling

To solve the issue of different sampling frequencies, we used upsampling to increase the frequency for the credit dataset. Downsampling would lose valuable information in an already sparse dataset.

Ridge Regression

Ridge regression is a linear regression model with coefficients estimated by ridge estimator. The coefficient in ridge regression is solved by the below optimization problem:

$$\hat{\beta} = augmin(y - X\beta)^T(y - X\beta) + \lambda(\beta^T\beta - c)$$
$$= (X^TX + \lambda I)^{-1}X^Ty$$

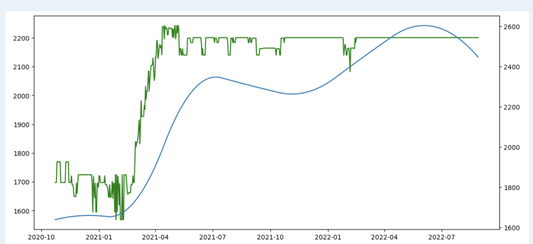
- Here, the independent variable is credit data and the dependent variable is the company's sales

XGBoost

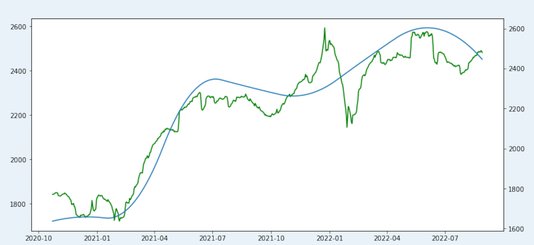
XGBoost is a library that implements ML gradient boosting algorithms efficiently, specifically tree boosting. Trees are built in parallel as opposed to sequentially.

Model Performance Results

Below we display the performance results of two prediction models:



(a) XGBoost



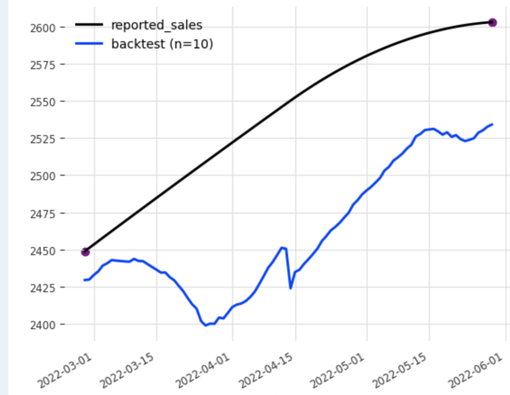
(b) Ridge Regression

- **Figure A (XGBoost):** Above graph shows the results yielded by implementing XGBoost on the denoised dataset.
- **Figure B (Ridge Regression):** The Ridge Regression model works very well – the regularization of coefficients decreases the variation and yields a stable model. This model yields an R^2 **score of 0.6192**.

Covariates

Future covariates are time series whose future values are known at the time of prediction, and often have known past values as well.

```
Backtest RMSE = 84.42869125176489
<matplotlib.collections.PathCollection at 0x7fba3af539d0>
```



We used credit data as future covariates in basic forecasting models and applied a simple Bayesian inference model to test the effectiveness of this approach, achieving **backtest RMSE 84.43%**.

Conclusion

- Ridge Regression model is best, with an R^2 **score of 0.6192**
- For this data, simple models are best - sample data is not sufficient for DL models to extract signals from noise

Enhancements

- Incorporate other alternative datasets to predict earning metrics, use general sector data as a predictor
- Supplement datasets and fill holes using simulation (e.g. GNN)
- Explore both old and new models after reworking data