Midterm Progress Report

Team IP1: Parker Jamison, Barbara Remmers, Sophia Su

Georgia Tech

Project Overview

Project Purpose and Introduction

This project aims to use specialized data from paper machine operations to predict a test on the short span compressive strength of containerboard[1]. This test is performed after a reel has been made and our business sponsors would like to predict the test in real time to have the opportunity to adjust operating conditions and enable loss reduction.

Goals

- An accurate model of compression strength test results using recent production data.
    - The compression test is the STFI Corrected Autoline Grade 93 test result (STFI)[1]
    - Model accuracy is measured by root mean squared errors (RMSE)
    - Modeling data is from one production line from 9/20/2021 -- 9/19/2022
- Using the model, identify critical aspects of the process and their impact on STFI test results

Team Progress

- Exploratory Data Analysis
- Data Wrangling and Feature Engineering
- Research on Cross-validation techniques and windowing
- Early Modeling

Exploratory Data Analysis (EDA)

NAs

Missing Values and Data Gaps

- The target value is missing from half of the observations.
- 16 numerical variables have over 80% missing values.
- There are time gaps in the data of 1-13 days. Grade Code has 10 missing values.

Erroneous Values

Erroneous Values and Mis-measurements

- The original dataset is not in chronological order.
- The sensors on the machinery may be error prone or reporting faulty values. (e.g., zero weights despite normal target values; a physical impossibility).

Others

Outliers, Categorical Features, and Other Data Characteristics

- Outliers are present in multiple features as measured by z-score, removing these records reduces the available training and test dataset considerably.
- Of the six categorical variables, three are effectively single valued, and two are redundant (index, grade code).
- The target variable is autocorrelated, which must be addressed in modeling.
- The process produces reels of one grade and then switches to another grade, often with a transition grade between two grades of very different nominal weights. This may be incorporated into modelling.

Data Wrangling and Feature Engineering

Data Issues and Resolution (Data Wrangling)

- Sorted data by Datetime index (critical for time series).
- Removed 25 single valued, redundant, or features with majority missing values as listed in Appendix I.
- Removed half of records without target variable.
- Further research required to determine best approach to address outliers in the data.

Feature Engineering

- Lasso and Elastic Net feature selection.
- Lagged Target Variable: A single lag is not enough to remove autocorrelation, per Durbin-Watson test. Not used.
- Lagged Target Variable 2: A weighted average of all earlier target values in the current batch of rolls with the same grade.
- Weight-based: The difference between the current reel's weight and a weighted average computed like Lagged Target Variable 2

Timeseries Cross-Validation Research

Research Paper Cited

- Cross validation for time-series research paper by Bergmeir and Benítez (2012)[2]

Techniques

- Blocked cross-validation
- Rolling Window/Walk-forward

Application of CV: Multiple Future Blocks for Testing

- Indicate useful lifespan of model before retraining is required
- The time boundaries of the blocks are set to changes in grade, to avoid the same information affecting training and test data (from backward looking features)

Modeling Approach: Bottom-Up and Top-Down

The **two directions** allows a valuable feedback loop:

- The **bottom-up** goal is to find a baseline multiple linear regression model
  - Small number of easily justified variables
  - Focus on feature engineering, which is easy to see the effects in this setting
  - Contribution to overall modelling effort
    - Provides a performance floor
    - Generates engineered features

- The **top-down** goal is to find a high-performing model: Lasso, Elastic Net, XGBoost, and Neural Networks
  - Considers all useable data
  - Benefits from technical aspects of complex models
  - Contribution to overall modelling effort
    - May provide higher accuracy to predict the target
    - Identifies additional important features for engineering

Models and Outcomes: Multiple Linear Regression

Explanation

- A straightforward model is easy to understand and justify implementation to business sponsors.
- Using a simpler modeling approach with some of the same features may bolster confidence in the more complex model
- The relatively simple setting eases feature engineering.

Outcome

- STFI is regressed on a weighted average of STFI lagged variable and a weight-based variable.
- Block CV with four folds estimates a preliminary RMSE of 1.45
- Note: Has not been tested against final holdout data.

Risks and Payoffs

- Risk: Ignoring important data and non-linear relationships.
- Payoff: Model predictions are better than using the last STFI to predict the current STFI (1.45 v. 1.6 RMSE).
- Payoff: Easy to interpret and justify to decision makers.
- Payoff: Easy to maintain and implement inline.

Models and Outcomes: Lasso and Elastic Net

Explanation

- Lasso regularization model used L1-norm to select the most important features.
- Elastic Net regularization model used L1-norm and L2-norm combined to select the most important features.
- L1-ratio = 0.7 for walk forward cross-validation
- L1-ratio = 0.8 for block cross-validation

Outcome

- Walk forward CV with four folds has a preliminary RMSE = 1.7.
- Block CV with four folds has a preliminary RMSE = 2.18
- Note: Has not been tested against final holdout data.

Risks and Payoffs

- Risk: Performance is largely affected by Cross-Validation method.
- Risk: Performance will be affected by tuning hyper-parameters.
- Payoff: Model is able to rank the importance of both numerical and categorical features.
- Payoff: Dimensionality reduction
- Payoff: Easy to interpret and justify to decision makers.

Georgia Tech

Models and Outcomes: Gradient Boosted Trees (XGBoost)

Explanation

- Ensemble methods like Gradient Boosted Trees may lead to higher accuracy.
- The XGBoost package can handle missing values.
- Tuning hyper-parameters is straight-forward.
- Provides some level of interpretability with feature importance scores.
- Beginners Guide to XGBoost.[3]

Outcome

- Initial modeling results are in the 1-2 RSME range on random test splits and before cross validation.
- Cross validation will recognize the temporal nature of the data. This will closely represent the production environment.
- This technique shows promise and will continue to be explored.

Risks and Payoffs

- Risk: Can lead to a high complexity model with "black box" properties, which hampers interpretability.
- Payoff: Increased prediction accuracy is possible after training on more data.
- Payoff: Feature importance scores can be used to reduce model complexity and provide business insights on what measurements are most important.

Plan of Activities

Activity	Status	Start	Target Completion Date
Exploratory Data Analysis	In Progress	05/19	06/15
Data Wrangling	In Progress	05/26	06/15
Feature Engineering	In Progress	05/26	06/15
Timeseries Cross Validation & Modeling Research	In Progress	05/26	06/15
Preparation of Mid-term Presentation	In Progress	06/02	06/15
Cross Validation Code Implementation	In Progress	06/02	06/22
Exploration of Model Methodologies	In Progress	05/19	07/01
Model Building & Validation	In Progress	05/19	07/07
Model Simplification & Dimensionally Reduction	Not Started	06/21	07/07
Analysis of Results	Not Started	06/21	07/07
Preparation of Final Paper & Submission to IP	Not Started	06/21	07/13
Final Paper Submission to GT	Not Started	07/13	07/19

Georgia Tech

Future Activities and Insights

- Models to Evaluate
  - Cross validate XGBoost and include engineered features.
  - Multiple Linear Regression – Further develop the engineered features (e.g., relative weightings of lags).
  - Further tune Lasso and Elastic Net for feature selection.
  - Potential to try Neural Networks

- Feature selection Business Subject Matter Experts (SMEs) techniques could suggest additional valuable explanatory variables.

- Model simplification and dimensionality reduction may slightly reduce model performance, but will lead to easier interpretation by business sponsors

- Analysis of CV results to understand model lifespan, retraining needs

References

1. TAPPI. (2013). *Short span compressive strength test of containerboard*. https://www.tappi.org/. Retrieved May 15, 2023, from https://www.tappi.org/content/tag/sarg/t826.pdf

2. Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

3. Seif, G. (2022, February 11). A Beginner's guide to XGBoost - Towards Data Science. *Medium*. https://towardsdatascience.com/a-beginners-guide-to-xgboost-87f5d4c30ed7

Georgia Tech

Thank you

Appendix I

Deleted Data Column Name	Data type	Reason for Deletion
DEFOAMER PP TO PM DAY TK | 29:0MSS050.PV	Categorical	Single valued
PM1 1 Wire/ 2 Saveall Defoamer Pump Stop/Run | 29:0MSS050B.PV	Categorical	Single valued
PM1 1/1A Saveall Defoamer Pump Stop/Run | 29:0MSS050A.PV	Categorical	Single valued except for one observation
GRADE CODE | 31:GRADE.STRING	Categorical	Redundant, use GRADE CODE | 31:ACTIVE_GRADE instead
INDEX | REEL ID NUMBER	Categorical	Redundant, use INDEX | DATETIME instead
Pine 1 Scan Residual EA | 27:0P1RESEA1.MN | Primary PLY	Numerical	Single valued
Pine 2 Scan Residual EA | 27:0P2RESEA2.MN | Primary PLY'	Numerical	Single valued
SAVEALL #1/1A DEFOAMER PUMP SPEED | 29:0HS050A	Numerical	Single valued
WIRE 1/ SAVEALL 2 DEFOAMER PUMP SPEED | 29:0HS050B	Numerical	Single valued
PM1 1st Press Air Bag Loading, Back | 4296	Numerical	More than 10,000 missing values

Appendix I (continued)

Deleted Data Column Name	Data type	Reason for Deletion
PM1 1st Press Air Bag Loading, Front | 4295	Numerical	More than 10,000 missing values
PM1 1st Press Panel Air Reading, Back | 4298	Numerical	More than 10,000 missing values
PM1 1st Press Panel Air Reading, Front | 4297	Numerical	More than 10,000 missing values
PM1 2nd Press Air Bag Loading, Back | 4300	Numerical	More than 10,000 missing values
PM1 2nd Press Air Bag Loading, Front | 4299	Numerical	More than 10,000 missing values
PM1 2nd Press Panel Air Reading, Back | 4302	Numerical	More than 10,000 missing values
PM1 2nd Press Panel Air Reading, Front | 4301	Numerical	More than 10,000 missing values
PM1 ENP Loading | 4305	Numerical	More than 10,000 missing values
PM1 Lump Breaker Roll Air Bag Loading, Back | 4304	Numerical	More than 10,000 missing values
PM1 Lump Breaker Roll Air Bag Loading, Front | 4303	Numerical	More than 10,000 missing values

Georgia Tech

Appendix I (continued)

Deleted Data Column Name	Data type	Reason for Deletion
PM1 Primary Freeness | 2951	Numerical	More than 10,000 missing values
PM1 Primary Headbox Consistency | 2950	Numerical	More than 10,000 missing values
PM1 Secondary Freeness | 2956	Numerical	More than 10,000 missing values
PM1 Secondary Headbox Consistency | 2948	Numerical	More than 10,000 missing values
REFINER #1 OUTLET PRESS | 29:1P1442	Numerical	More than 10,000 missing values