Lasso Cross Validation Model for Retail Overtime Prediction

Dominique R. Grimes

June 3, 2023

Lasso Cross Validation Model for Retail Overtime Prediction

Much of the work industry relies heavily on the productivity of their teammates. Workplace productivity is an attempt at measuring how teammates accomplish the production of quality goods and services in alignment with the organization's strategic goals and objectives. The productivity of employees is essential to a company's overall success in an ever-increasing completive market. Often businesses require employees to work overtime to reach target productivity measures. The focus of this model is to estimate the amount of overtime needed per worker to reach a certain level of productivity.

Overtime is expensive and can cause employee burnout. Businesses should balance this requirement with caution since too much overtime over extended periods of time can have other adverse and costly impacts to the organization. It is valuable to an organization to understand the cost in overtime compared to the incremental output of productivity gained. It is also essential to monitor if that level of overtime is sustainable long-term.

Top organizations are looking for creative ways to do more with less resources, expense, time, and employee turnover. There are several internal and external factors that may impact an individual's productivity. These items can vary from work or home environment, team culture, amount of time worked, if the person enjoys what they're doing, if they are enrolled in school or have children, etc. Other market competitors are getting creative to attract top talent, including flexible benefits and work-life balance.

The Data

Data Description

The dataset I'm using to create the model was obtained on Kaggle. This data is based on one organization's clothing retail production, which is heavily dependent on human labor. Gaps in predicted production versus actual production can cause major issues in clothing supply and demand. Each row represents data based on daily metrics. There are twenty-six columns in the original data file. They are team, targeted productivity, allocated time for a task

(svm), number of unfinished items for products (wip), overtime, incentive, idle time, idle men, number of style change, number of workers, month, quarter 1, quarter 2, quarter 3, quarter 4, quarter 5, finishing department, sewing department, Monday, Tuesday, Wednesday, Thursday, Saturday, and Sunday.

Data Cleansing

The data requires minimal cleansing. I am combining the two department finishing columns into one, as there was a space at the end of one of the column headers. I am correcting the spelling on the "department_sweing' column to "department_sewing". Some of the variables are already separated into dummy variables. The team and month also need to be represented as dummy variables in preparation for modeling.

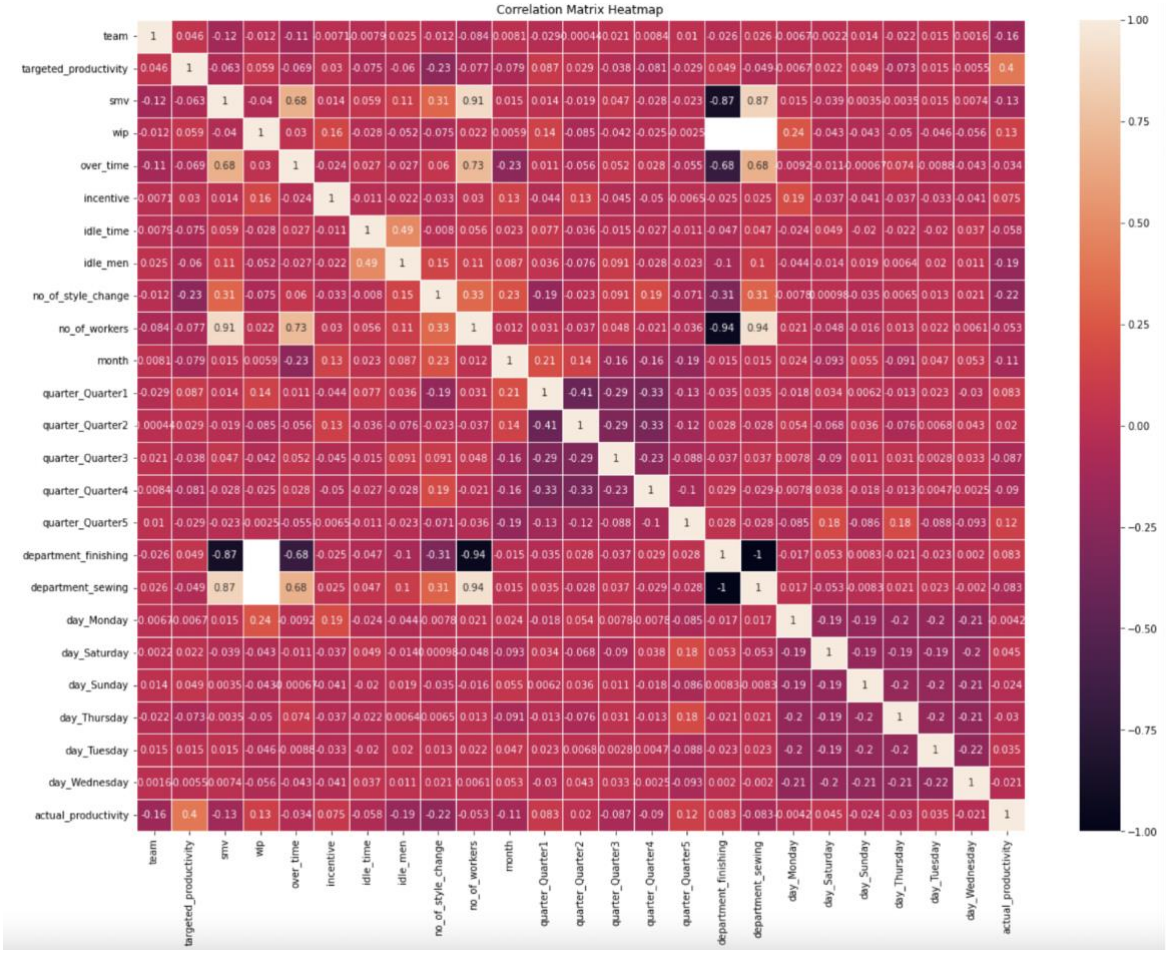
I am also managing missing values in the data by observing any columns with over forty percent missing values. This criterion resulted in the 'wip' column needing further investigation. Analysis showed that the null values were all coming from the finishing department. Since the values are not random and coming from one department only, I'm removing this column to avoid bias in the model. In addition, the correlation between 'wip' and the target variable is low.

For the target variable, I am creating over_time/no_of_workers as a new feature for the which divides the values in the over_time column by the no_of_workers column. This is being created since the total overtime in each row is dependent on the total number of workers for that team.

Exploratory Data Analysis

My graphical analysis consists of a correlation heatmap, scatter plots, and a bar chart. In the correlation heatmap department_sewing, no_of_workers, and svm have strong positive correlations with over_time, .68, .73, and .68 respectively. The strongest positive correlations overall are department_sewing (.94), number_of_workers and svm (.91), and department_sewing and svm (.87). These correlations lead to further graphical analysis through

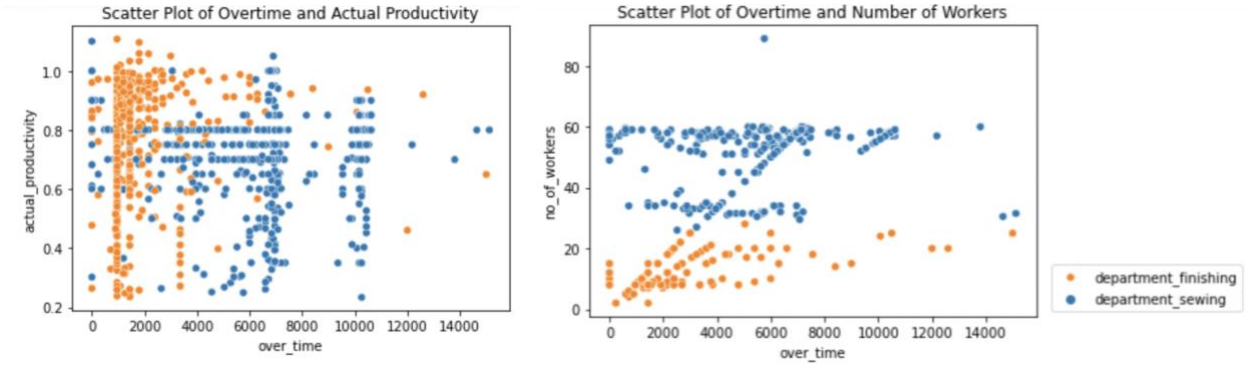
scatter plots. I took a further look at the relationships between the highest correlated values with over_time. In addition, I viewed a scatter plot of actual_productivity and over_time.



The first scatter plot shows over_time versus actual_productivity and is categorized by department. From the coloring of the plot, department_finishing has less overtime than the department_sewing. The lower hours in overtime tend to have higher concentrations in productivity, whereas the higher values of overtime have moderate productivity.

The next plot shows over_time by no_of_workers, categorized by department. From the coloring of the plot, the finishing department has significantly less workers than the sewing department. There are a few linear relationships emerging from the graphical analysis. This leads to

viewing the categories by teams in lieu of departments; however, no prevalent patterns were observed.



One additional relationship analysis is focused on the correlation between svm and actual_productivity. The finishing department has consistently low svms with low variability, and the sewing department has much more variability with higher svms. Finally, I'm looking at the distribution of row data entries by team and department to get a better idea of the distribution of data. Each sewing team has a similar volume of row entries with an approximate range of five, while the finishing department has more variation in the total metrics reported by each team with a range of approximately twenty records. The sewing department has higher counts than the finishing department overall.

The Model

Preprocessing, Hyperparameter Tuning, and Grid Search Cross Validation

I am splitting the data into 80% training, 10% validation, and 10% test. Since the target variable is continuous, I am fitting regression models for comparison. The three models for comparison are Polynomial Regression, Ridge Cross Validation Regression, and Lasso Cross Validation Regression. I am using StandardScaler which is generally the industry go to for scaling. Regression models tend to benefit from normally distributed features, and standard scaler adjusts the mean to zero and creates a standard measure of variance. For feature selection, Grid Search Cross Validation is my chosen method in lieu of methods such as

VarianceThresholder or PCA. I'm also hyperparameter tuning the models with degrees of one, two, three, and four for the polynomial regression model and regressor alphas of one-tenth, one, and five for both Ridge and Lasso regression models. A five-fold cross validation is being used in the Grid Search.

Model Selection and Evaluation

I am using r2 score and root mean squared error (rmse) to evaluate the fit of each model, which are common evaluation metrics for regression models. R2 is a great comparison tool when evaluating different regression models against each other, as it is measured independent of context. R2 shows the percentage of variability that the model accounts for. The rmse can be used as a loss function and is sensitive to outliers. We want to account for the maximum amount of variability and minimize the loss in the model performance.

The best model produced from the Grid Search Cross Validation is a Lasso Regression, with Standard Scaler, a polynomial degree of two, one thousand max iterations, no alphas, n alphas of one hundred, and tolerance of ten-thousandths. When fitting the model, I received some convergence errors. Despite this, the validation set r^2 value is calculated as .90 and the rmse is 17.5.

Once concern is overfitting the model. I am accounting for this by using a test data set on the model. The test data applied to the model resulted in improved r2 over the validation set with an r2 of .93 and consistent rmse of 18.8.

Conclusion

Summary of Findings

The graphical analysis shows a few features that are correlated with overtime. There are two departments and twelve teams. Overall, department_finishing has fewer teammates, lower svms, less overtime, and higher productivity than department_sewing. Polynomial Regression was a starting point for fitting the model; however, Grid Search Cross Validation with LassoCrossValidation and RidgeCrossValidation Regression resulted in a model that accounts

for over 90% of the variability of the data with an estimate around seventeen minutes of overtime per person required to meet the target productivity.

Considerations

There are items to consider before deployment. Quality assurance testing will need to be completed and the Convergence Warnings need consideration on if they can or should be cleared. One challenge of fitting the model is the processing time of a couple hours to fit the model. The alpha value can be optimized as there were no features eliminated in the best model. An additional opportunity is to explore are Elastic Net Regression. This model may be more efficient than the Lasso Regression Cross Validation. There is also an opportunity to analyze outliers and how those values are impacting model performance.

Recommended Use Cases

Overall, the objective to create a model that estimates the amount of overtime required per teammate based on several features is effective. Practically, this tool can be used as a decision-making tool to either pay overtime, extend a deadline, or justify hiring more workers. As a long-term strategic tool this model may be used as a foundational tool to measure the positive or negative impact that overtime has in relation to other variables on organizational success, most importantly the wellbeing of the workers that contribute to the end goal.