Optimizing Protein Titer Production with Chinese Hamster Ovarian Cells: Predictive Modeling and Recommendations for Enhanced Yield

Jackson E. Polk

Summer Semester, 2023

DSA 5900

4 Credit Hours

Supervised by Dr. Talayeh Razzaghi

# Introduction

Cytovance Biologics (Cytovance), a biomanufacturing company located in Oklahoma City, is seeking analysis to maximize the production of recombinant protein concentration, or titer, from animal cells. Their Chinese Hamster Ovary (CHO) cell line is utilized in preclinical to early-phase of therapeutic protein development [2].

The production of recombinant proteins generally involves introducing an “expression vector” into a “host” organism, which will then use the instructions contained in the vector to synthesize and reproduce more of the targeted protein [6]. The host is usually a cell from either bacterium, yeast, an insect, or a mammalian cell culture [6]. With their ability to complete proper protein folding, mammalian cells remove an additional manufacturing step when compared to bacteria and yeast cells [7]. Additionally, mammalian cells excel in glycosylation, a critical process in several biological processes [1]. However, their longer production time necessitates maximizing output efficiency. Attempts have been made to identify crucial factors in CHO cell production, with a “temperature shift” strategy during production being proposed recently [8]. Regardless, Cytovance attempted their own experiment with different inputs.

Cytovance experimented with several different input combinations across 13 input variables, attempting to determine which combination produced the highest concentration of these recombinant proteins. Each experimental duration ranged between 3 and 14 days. With these results, Cytovance would like a machine-learning model to be developed to aid in the study of how these inputs can be tuned to produce optimal protein titer. They have sought help from the University of Oklahoma to produce a model and allow this analysis to be done.

## 1.1 Objectives

To facilitate faster and more cost-effective analysis of protein titer production, this project focuses on developing a machine-learning model capable of accurately capturing titer behavior. The project aims to achieve two main goals:

1. Model Development: Create a machine learning model from the provided data that will adequately capture titer by octet, or titer, production behavior. The best model will be identified by the Root Mean Square Error (RMSE), evaluating its predictions against the actual titer.
2. Statistical Analysis: Provide a thorough analysis of the developed model that will infer relationships between the input and output variables. This will identify key factors the influence successful titer production.

Throughout the project, several other challenges will be addressed intelligently, particularly surrounding data quality. The provided dataset totals around 2000 observations, and most are missing observations for the outcome variable, titer. Therefore, missing values will need to be imputed using an unbiased imputation method. Additionally, an un-biased training set will need to be created. In summary, this project aims to develop a machine-learning model for protein titer analysis, determine relationships between input and output variables, and address various challenges to enhance the efficiency and accuracy of protein production.

# Data

The experimental data from Cytovance has 2076 records with 13 input variables and 18 output variables including the target output, titer. The data describes several experiments conducted by Cytovance, each using unique combinations of input variables to observe a change in titer production level. The raw data is in the form of two Microsoft *Excel* files, one for each CHO strain (KC, S). The CHO-KC file contains 3 data sheets with a total of 896 records, and the CHO-S file contains 8 data sheets with a total of 1180 records.

Each experiment is uniquely identified by its corresponding CHO Strain (KC or S), the *Excel* Sheet, vessel type and vessel name. The experimental period ranged anywhere from 3-14 production days. A major challenge faced by Cytovance was that data collection disturbed the titer production process, leaving several missing observations of titer. To solve this, Cytovance opted to record titer every other day. A breakdown of the missing values is shown in Table 1.

Table 1 Quality of Raw Data Provided by Cytovance



The two main challenges for addressing are the small sample size and the large number of missing values. Prior to addressing these challenges, these data will need to be combined. This transformation step can be done using Microsoft *Power Query* by applying the following transformations:

1. Remove any rows not directly associated with the main data table.
2. Remove any unnecessary blank columns.
3. Append all data together, keeping note of the *Excel* sheet the data came from.
4. Transform character variables to lower case.
5. Transform values reported as “N/A” to *null*.
6. Remove special characters (from columns such as “Temp” and “pH Setpoint”).
7. Perform obvious data type transformations (e.g., “Temp” as numeric).
8. Omit completely *null* rows, especially if missing “Vessel Type.”
9. Create numeric column “Vessel Volume” using information from the slides given by Cytovance.

## 2.1 Quality

### 2.1.1 Input Variables

From the input data quality report, new challenges are identified. There are large magnitude differences (see target cell seeding density in Table 2), there are a low number of unique values within the numeric inputs and some variables with little variance.

Table 2 Input Variable Data Quality Report



*Note:* Variable names with asterisks (\*) by name were the result of combining data step.

From the ingestion process, three new variables are added. Of these, vessel volume has a one-to-one relationship with vessel type, as each vessel type only had one unique vessel volume.

### 2.1.2 Output Variables

Within the data, there are seventeen output variables in addition to titer. The descriptive statistics of these can be seen in Table 3.

Table 3 Output Variable Data Quality Report



Notice that the variable of interest, titer has more missing values (67%) than other outputs. However, viable cell density has about 8% missing values. There are also correlations amongst these variables, which will be useful for recovering missing observations of titer. The correlation plot matrix is given in Figure 1. These relationships lend themselves to recover titer using “Multiple Imputation by Chained Equations” (MICE) [5].

## 2.2 Exploration

To implement MICE, some of the other 17 output variables must display suitable relationships with titer. A visualization of some these relationships is shown in Figure 1.

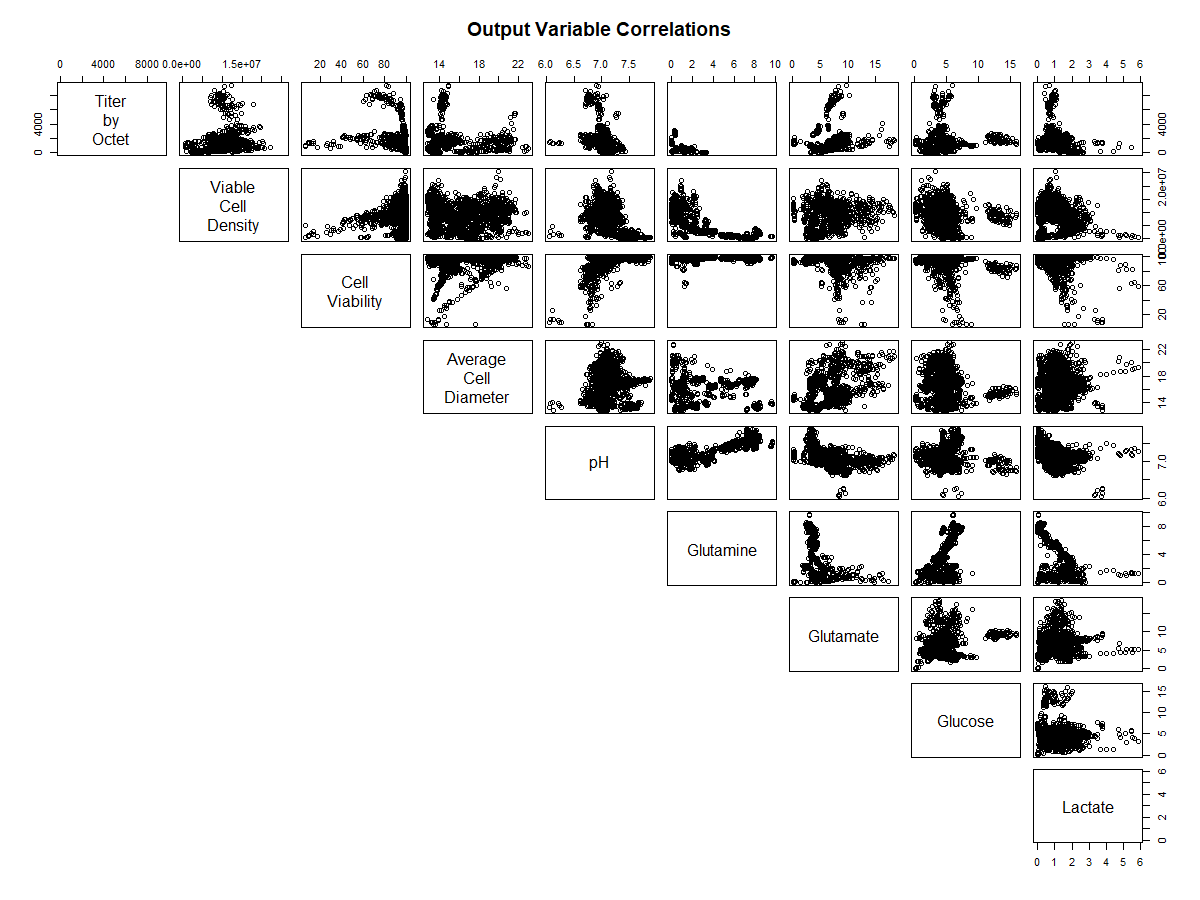


Figure 1 Correlation Plot Matrix for Output Variables

The outputs with the highest absolute correlation to titer are glutamine, pH, cell viability and osmolality. Some of the output variable relationships are extraordinarily complex and non-linear. To resolve this issue, MICE has options for models that even capture non-linear relationships, such as random forest regression.

Titer has a strong relationship with production day, which is a time element. This is in line with previous knowledge of titer, given that mammalian cells take longer to develop adequate production levels.

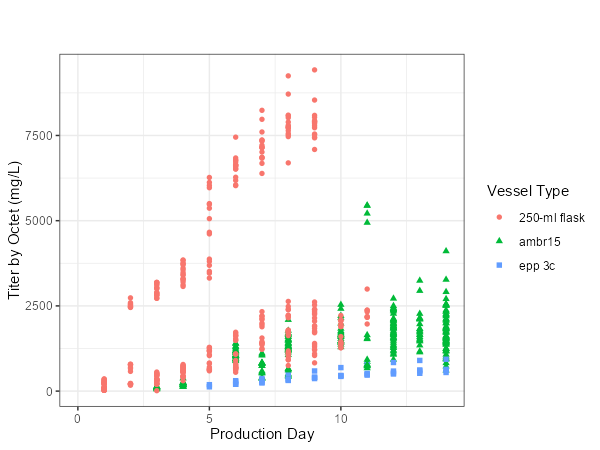


Figure 2 Titer vs Production Day

There appears to be an interaction effect between the vessel type and the rate of titer production. From Figure 2, the 250-mL flask titer production behavior has a subset of titer records with a quite different pattern. This relationship should be investigated further within this data exploration phase.

Table 4 Unique Values in Input Variables



Table 4 shows how that there are a small number of unique values in the input variables associated with Cytovance’s ‘250-ml Flask’ Vessel Type, particularly when compared with the ‘AMBR-15’ and ‘EPP 3c’ vessels (Table 4). This, in combination with the 250-mL Flask’s differing interaction with production day (Figure 2) warrants the creation of a separate model for this case. Therefore, the observations with ‘250-mL Flask’ will be separated from the traditional modeling data. This leaves 440 observations for the ‘250-mL Flask’, and 1636 observations to be modeled separately.

Figure 3 displays a strong relationship between titer, production day, and target cell seeding density within the 250-mL flask subset of data.

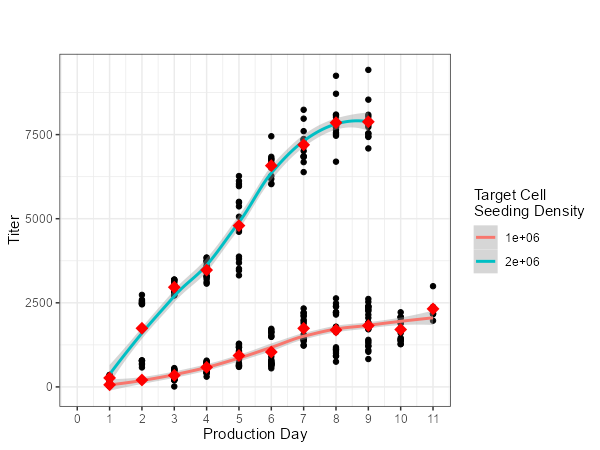


Figure 3 250-mL Flask Titer vs Production Day

*Note:* Average shown with red points, lines shown are smoothed Loess curves.

Given that these variables are major predictors, an attractive modeling approach may be to create an independent Time Series ARIMA model for each target cell seeding density level and observe how the forecasts of this model perform.

## 2.3 Pre-Processing

The objectives for pre-processing are to recover as many of the missing titer observations as possible, scale the numeric inputs, and convert categorical variables to numeric using “One-Hot” encoding. These steps will curate the data properly for the variety of machine learning models that will be applied.

### 2.3.1 Data Imputation

Prior to applying predictive models to titer, recovering as many of the missing observations as possible will potentially enhance model performance.

A graph with a line and a line graph

Description automatically generated

Figure 4 Example of Missing Data within Titer

Figure 4 shows how the missing data occurs at distinct intervals. For this reason, imputation by time series (linear, spline) may be attempted in addition to the MICE imputation methods to recover the missing datapoints. Excluding the ‘250-mL Flask’ data, 79% of the records are missing a titer value. It is crucial that the bias toward imputation method be addressed. Therefore, several imputation methods will be tried and compared to training on the unimputed data:

1. MICE
   1. Classification and Regression Trees (CART)
   2. Least Absolute Shrinkage and Selection Operator (LASSO)
   3. Random Sample
   4. Random Forest Regression
2. Time Series Interpolation
   1. Linear
   2. Spline

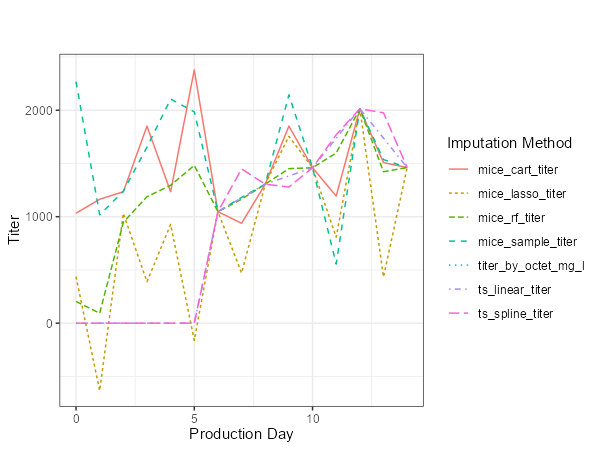


Figure 5 Example of Imputation Method Behavior

The key difference between the MICE and Time Series methods is the unpredictable behavior, best displayed in Figure 5. Another difference is the behavior of titer before the first “non-zero” observation. Importantly, Cytovance noted that titer on production days prior to the first non-zero observation may not necessarily have zero titer produced, which is consistent with the MICE imputation results. The Time Series methods impute these with 0, where MICE use the other output variables to impute this value of titer.

### 2.3.2 Scaling and Encoding

To handle the scale differences in the input data (notably between target cell seeding density and pH setpoint) each numeric variable was mean-centered and scaled. There were few distinct categories within the categorical variables which meant ‘dummy variables’ using “One-Hot” encoding would be an appropriate option.

### 2.3.3 Outliers

Amongst the numeric input variables, there were only a small number of outliers (see Table 5). This was determined by checking which values were greater than three standard deviations above the attribute average. None of these outliers are expected to affect model performance.

Table 5 Outliers Within Numeric Inputs



*Note:* all 60 Vessel Volume outliers are attributed to the ‘EPP 3c’ Vessel Type, which has a volume of 3.75 L.

# 3. Methods

To analyze which model and imputation method results in the best performance, the independent variables of imputation method, model type and tuning parameters can be changed to observe the RMSE against actual titer. Within the Data Exploration phase, it was determined that separate model types would be necessary for the 250-mL flask data subset. For this, two separate ARIMA models will be developed to represent how the level of Target Cell Seeding Density affects the titer output. The remaining data will be analyzed using traditional predictive modeling techniques.

## 3.1 250-mL Flask Time Series

One of the major components of titer production is time, and because there are many missing or incomplete cases within the 250-mL flask subset of the data, time series methods can be used. Specifically, an Autoregressive Integrated Moving Average (ARIMA) model will be used for its versatility in capturing trend. To create a dataset that was adequate for time series modeling, titer values in this subset were averaged by day and their corresponding Target Cell Seeding Density value, exactly as seen in Figure 3. This resulted in one data point per unit time for each of the models to be trained on. For each level of Target Cell Seeding Density, the final 3 Production Days’ values were held out to test the model’s forecast RMSE.

### 3.1.1 ARIMA Model

To determine an acceptable ARIMA model (1), there are 3 parameters to tune. These are the number of autoregressive (P) and moving average terms (Q), as well as the differencing (D).

(1)

This model will determine the future values of titer by only considering previous values of titer in various capacities. The term represents the differenced time series (the current value of titer) differenced by D values before. The terms are weights on the autoregressive components, and are weights on the moving average, or smoothing component. Tuning these models follows the grid search method, where each combination of terms (P, Q, and D) will be modeled and tested independently. The combination that produces the lowest forecast RMSE will be chosen.

## 3.2 Regression Models

### 3.2.1 Test Set Creation

Due to the low number of datapoints available for training, creating a test set will need to be done with care and with as little bias as possible. 20% of the data was held for model testing, with the remaining data to be split between training (80%) and validation (20%). This resulted in 328 observations within the test set, 1048 in the training set and 260 in the validation set.

### 3.2.2 Predictive Models for Titer Level

To protect against model selection bias, five separate machine learning models will be investigated in this experiment. In the data exploration phase, several non-linear relationships were identified between titer and its inputs. However, models that capture non-linear relationships are more difficult to interpret. For this reason, there are two models best suited for linear relationships (linear model and elastic net), two models for non-linear relationships (random forest regression, gradient boosted regression trees) and one median model type in Multivariate Adaptive Regression Splines.

#### Multivariate Adaptive Regression Splines

The major advantages of Multivariate Adaptive Regression Splines (MARS) is that it is can approximate non-linear relationships, perform feature selection, and determine important interactions between variables. This is accomplished by creating piecewise-linear regression functions. The algorithm will determine several hinges within the linear model which will then be “pruned” if the resulting piece-wise model does not generalize well to new data. This same process also provides a framework for feature selection, which is a benefit to the objective of determining important inputs in predicting titer.

#### Gradient Boosted Regression Trees

Gradient Boosted Regression Trees (“GBRT” or “GBM”) is a popular algorithm that can capture non-linear relationships within data. It accomplishes this by sequentially adding predictors and allowing them to “correct” their predecessors [3]. The error left over by the predictors predecessors becomes the target variable in which a new decision tree regression is trained.

#### Random Forest Regression

Random Forest Regression is an ensemble method like GBRT, which uses the power of aggregating multiple predictions to produce a better prediction. For instance, two trees in a “forest” can produce powerful predictions of titer if they come about these predictions independently. Within the context of titer prediction, predictions from one tree that emphasizes production day and dissolved oxygen and predictions from another tree that emphasizes target cell seeding density and temperature will produce better results. Random Forest Regression will create many of these forests and aggregate their predictions and yield an especially useful “importance” metric for each input variable.

#### Linear Model

The Linear Model is the simplest model and its success in predicting titer would give a straightforward answer to the objective of analyzing the relationship between the inputs and titer. For instance, the linear model will train coefficients as seen in (2) that will represent the simple gain by increasing the input variable by a single unit.

(2)

For instance, if a coefficient corresponding to temperature were produced, it would be simple to communicate that increasing temperature by a single degree produces an increase in titer by two cells per milliliter.

#### Elastic Net

While Elastic Net is a modeling technique, the true power is in the combination of two coefficient regularization methods (LASSO and Ridge Regression). It utilizes two parameters, and to penalize model coefficients. The parameter regulates the total penalty, in which a value of zero would incur no penalty and result in the simple linear regression model. The mixing parameter, determines the emphasis placed on LASSO coefficients over Ridge Coefficients. A higher produces coefficients that are closer to the results of LASSO.

(3)

A major advantage of using Elastic Net is that it allows for feature selection. As an example, variables that have no relationship with titer will have a corresponding coefficient of zero, indicating zero importance.

Each of these models has a degree of tuning, which will be used to help generalize the training data to the test data. A breakdown of the model, tuning parameters, and the values that will be tried are shown in Table 6.

Table 6 Machine Learning Models and Tuning Parameters



*Note:* Random Forest “Split Rule” parameter ‘variance’ refers to reducing the variance at the leaf node, ‘extratrees’ refers to extremely randomized trees, and ‘maxstat’ refers to the maximization of certain statistics.

For the business objective, providing a model that allows insights into interactions between the input variables and titer is incredibly important. Each model was trained using 5-fold cross-validation with two repeats. This will provide a robust training procedure which ensured low bias of the model toward the training set that was selected.

## 3.3 Variable Importance

To complete the business objective of describing the important factors in titer production, the concept of “variable importance” will be used. The exact implementation follows the caret::varImp function [4]. For each model, there is a different method of calculating variable importance.

#### Linear Models

Linear Models report feature importance by using the absolute value of the *t-*statistic for each model parameter [4]. The *t-*statistic in linear regression is a proxy for the significance of a feature’s individual coefficient. For instance, some features may warrant a very large coefficient but be deemed insignificant predictors by the *t-*statistic.

#### Random Forest/Boosted Trees

Forest and tree-based models utilize a permutation style method, assessing the reduction in RMSE by adding a variable of interest [4]. This is done first at the individual tree level, and then summed across an entire forest. For example, this method would evaluate the reduction in titer prediction error when the forest adds production day as a predictor.

#### MARS

MARS reports importance in a similar fashion to forest and tree-based methods; it is the reduction in cross-validated statistic (such as RMSE) by addition of the variable of interest. The important distinction between MARS’s importance and tree-based methods is the scale of the importance metric itself.

Utilizing the importance metrics will help identify a smaller subset of input variables that can be used in an analysis of variance.

## 3.4 ANOVA on Important Features

Once important features have been identified, their interactions can be analyzed using an analysis of variance, or ANOVA. Each variable will be assessed for its singular effect on titer, as well as its interaction with the other variables. It is vital to understand that this ANOVA will observe the effect on the imputed titer data to provide as many complete records for analysis as possible. The imputation method of choice will be determined using the imputation method that produces the models with the lowest average test RMSE. The intention here is to create a small subset of variables and interactions that can be assessed visually and presented to Cytovance.

# 4. Results and Analysis

## 4.1 250-mL Flask Time Series

The grid search method was used to find the ARIMA’s hyperparameters. Various hyperparameters were used for ARIMA model, and two sets of results were obtained for the 1 million and 2 million levels of Target Cell Seeding Density. At the 1 million cell level, the best ARIMA model with parameters P = 1, Q = 1 and D = 1 produced an RMSE of 223.19. At the 2 million cell level, the best ARIMA model had parameters P = 2, Q = 2 and D = 0 produced an RMSE of 314.61.

A graph of different colored lines

Description automatically generatedA graph of different types of data

Description automatically generated

Figure 6 Grid Search Results for ARIMA Model Parameters

Figure 6 shows how test RMSE decreases as more autoregressive terms ( from Equation 1) are added in the two million cell level. This indicates that future titer production is highly correlated to previous titer production. At the one million cell level, there was little consistency in how tuning parameters affected the results.

A graph of a graph showing the growth of a cell seeding

Description automatically generated

Figure 7 ARIMA Model Forecast Performance

Both ARIMA model forecasts and their tuning parameters provide insight into how titer production behaves as a function of time, as shown in Figure 7. The major indications from the models are that target cell seeding density is a major factor effect on titer production, and that in the two million target cell seeding density case there is much more of an influence of previous titer production, as indicated by the two autoregressive terms. The forecast in Figure 7 also shows a possible taper in production at around 8000 titer in the two million target cell seeding experiment.

## 4.2 Predictive Models

As displayed by Table 7, the experiment found that imputation method had a significant effect (p < 0.05) on the resulting RMSE. Between the sampling methods, it was found that time series imputation methods, both linear and spline, performed worse in the creation of a training set. In fact, those methods were among three that had an average test RMSE than training with no titer imputation at all.

Table 7 Imputation Method Effect on Average Test RMSE



According to a Tukey comparison of means just two imputation methods, MICE with LASSO and MICE with Random Forest performed significantly better (p < 0.05) than no imputation at all.

The experiment also found that model type had a significant effect (p < 0.05) on the resulting RMSE. Each model performed significantly different than the other (p < 0.05), with two exceptions being the differences between GBRT and Random Forest, as well as Linear Modeling and Elastic Net. However, it will still be important to check the interaction effect between model type and imputation method. Table 8 shows the model type and imputation method that produced the lowest average RMSE on the test set.

Table 8 Best Imputation Method by Model Type



Table 10 shows that, in terms of average RMSE the best combination of model type and imputation method is a Linear Model trained using data imputed by MICE with LASSO. This is convenient for model interpretation, however there may be variance within the test results as seen in Figure 8.

A graph with a diagram

Description automatically generated

Figure 8 Model Combination Performance Comparison

A pairwise comparison reveals the linear model does outperform the only significant difference (p < 0.05) in model type combination is between gradient boosted trees and each of the other model types. None of the other model types differ from each other, statistically.

### 4.2.1 Hyperparameter Results

Analyzing the hyperparameter tuning results from each model types’ best results can provide valuable insight into how algorithms interpret this data. Table 9 displays each models tuning parameters. The linear model having no intercept correctly interprets that, with a value of zero for every factor, notably production day, the value of titer should be zero. The MARS model degree of 2.55 shows that there is, on average, two or three level interaction effects that produce meaningful results within titer. Elastic Net’s parameter “Alpha” shares that there should be an even mix between LASSO and Ridge penalties, and “Lambda” shows that there are definitely penalties needed for this model. From the tree-based methods, the interaction depth from GBRT and MTry from random forest are particularly interesting for analyzing how each of the variables affect the other. These methods show that generally there are several interactions within the data that capture variance within expected titer production.

Table 9 Average or Most Frequent Model Hyperparameter Results



## 4.3 Variable Importance

An overall feature ranking can now be created using the variable importance results. Since Gradient Boosted Trees performed objectively worse than the other model types, its variable importance rankings will not be taken into account. Each model produces importance values that are on different scales, and therefore a simple rank-average system should be used. Table 10 presents these results.

Table 10 Variable Importance Ranks by Model Type



From these results, it is evident that the top five features in describing titer production are production day, glucose trigger limit, temperature, media type, target cell seeding density. An analysis of variance (ANOVA) procedure can be used to analyze how these variables impact titer production.

### 4.3.1 ANOVA on Important Features

The results of applying ANOVA on the key features is shown in Table 12. Note that only the statistically significant results are displayed.

Table 11 ANOVA Significance of Important Variable Interactions



From Table 11, Media Type on its own is insignificant, however its interaction with both time and temperature are important and worth further consideration. Additionally, temperature along with its interaction with time corroborate the results of previous literature [8].

#### Singular Factors

Production Day

Titer production was previously known to increase with time [6], and the results of this experiment corroborate that idea (see Figure 2, Table 11).

Glucose Trigger Limit

Increasing the glucose trigger limit from a lower value to a higher value can increase expected titer production, as shown by the average titer value displayed in Table 12.

Table 12 Glucose Trigger Limit Effect on Average Titer



Target Cell Seeding Density

Target cell seeding density was identified as a major factor in the time series model, and the same is true within the predictive model subsets. From Table 13, increasing target cell seeding density can increase expected titer production.

Table 13 Target Cell Seeding Density Effect on Average Titer



Temperature

Temperature has been known to influence titer [7], and the results of this experiment show no difference. Increasing the temperature is shown to increase the expected titer production, as seen in Table 14.

Table 14 Temperature Effect on Average Titer

#### 

#### Factor Effects Over Time

Many of these individual effects have interaction effects with production day. In other words, their interaction with titer changes as the given experiment length increases. This is particularly true for temperature.

Table 15 Interaction Effect Between Temperature and Time



Table 15shows that beginning at a lower temperature and shifting to a higher temperature in later production days may be ideal. This reiterates the dynamic relationship that titer has with temperature and time. Time also has an interaction with Media Type. However, this relationship is described in a unique way, because media type cannot be changed within an experiment. Instead, Table 16 shows that experiments that are shorter should prefer media type B.

Table 16 Interaction Between Time and Media Type



Interestingly, media type B and glucose trigger limit produce the exact same results, where a high glucose trigger limit is preferred for shorter experimental runs but does not matter with longer-term production. These results are shown in Table 17.

Table 17 Interaction Between Time and Glucose Trigger Limit



The interaction between temperature and media type is the final effect worth investigating. Although there were no observations for titer with an elevated temperature and media type of B, the interaction effect is seen at the lower temperatures.

Table 18 Interaction Between Temperature and Media Type



Table 18 shows that at low temperatures, media type B outperforms other media dramatically.

# 5. Conclusion

From this project, the most accurate model for predicting titer was identified, and all other modeling results were used to aid in identifying relevant methods for increasing titer production. From imputing missing data, training to a variety of model types, analyzing importance metrics and finally ANOVA, a robust solution to Cytovance’s question can be proposed.

## 5.1 Final Model

There were three models produced for Cytovance. There were two independent time series models based on the 250-mL Flask Data, one for each level of Target Cell Seeding Density. These models produced forecasts with RMSE values of 223.19 at the 1 million level and 314.61 at the 2 million cell level. Finally, the best-performing model was the linear model, trained on data imputed using the MICE with LASSO. This yielded a RMSE of 1117. Investigating all of the different modeling methods used, their hyperparameters, and variable importance metrics provided us with useful suggestions to present to Cytovance.

## 5.2 Deliverables to Cytovance

From the models that were created and based on the data provided, the following suggestions can be made:

**Case 1: Vessel Type - 250-mL Flask**

Set the Target Cell Seeding Density to two million cells/mL and cap production time at 9 days. Following this period, the ARIMA model projects that titer production will see diminishing returns.

**Case 2: Vessel Type – AMBR-15 or EPP 3c**

According to the results of section 4.3, the most important factors in titer production are time, temperature, target cell seeding density, media type, and the glucose trigger limit. Each one of these factors produces significant changes in the expected titer production. Titer production is obviously affected by time, however time also affects how the other factors interact with production. The only factor not affected by time is target cell seeding density, in which the higher this is, the better.

The intended production length matters to selections of media type, glucose trigger limit, and temperature. For instance, shorter production runs should utilize a high glucose trigger limit, media type B, and lower temperatures. However, for longer production runs, glucose trigger limit and media type matter less, but the temperature should be shifted higher in the latter half of production.

## 5.3 Future Directions

There were numerous aspects that were excluded from final analysis due to either time or resource constraints. These include but are not limited to tuning imputation methods, such as performing hyperparameter tuning on the MICE imputation method that utilizes regression trees (CART). Additionally, more model types could be attempted, incorporating neural networks. Future models should explore the non-linear relationships within this data, and model choices should be made accordingly. Finally, an extended analysis of input variable that identifies firm values rather than the general guidelines produced in this project would be of great value to Cytovance.

# 6. Self-Assessment

Throughout the project, I progressed significantly as a professional data scientist. There were many challenges and frustruations, but these were alleviated with perseverance and an open mind to learning. The most important lesson from the entire project was having the patience to stay organized. There were several points during this project that required re-working, re-factoring, or validation. However, early in this project I had rushed through the process and lost some of the work that had already been done. Had I maintained an organized structure, particularly on GitHub and in my local files, I may have performed significantly better on this project. Another lesson learned was communication. I had an excellent advisor in Dr. Talayeh Razzaghi, and her help on this project proved invaluable. We met twice a week to go over results, and this proved vital to finishing this project on time. I acknowledge and respect that others, particularly those who are as well-versed as Dr. Razzaghi, can play in the success of a project.

References

[1] Biology Lectures. (2021). *Glycobiology | Glycosylation of proteins | Factors affecting glycosylation |*. *YouTube*. Retrieved July 16, 2023, from https://www.youtube.com/watch?v=qFqbueKBd8w.

[2] Cytovance Biologics (n.d.). *Mammalian Biologics CDMO Services.* Cytovance. <https://cytovance.com/cdmo/mammalian/>

[3] Geiron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems (2nd ed.). O'Reilly.).

[4] Kuhn, M. (2023). caret: Classification and Regression Training [Software]. R Foundation for Statistical Computing, Vienna, Austria. [https://CRAN.R-project.org/package=caret](https://cran.r-project.org/package=caret)

[5] Raghunathan, Trivellore & Lepkowski, James & Hoewyk, John & Solenberger, Peter. (2000). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Survey Methodology. 27.

[6] Recombinant Protein Production. (2021, December 16). City Tech CUNY. https://bio.libretexts.org/@go/page/75276

[7] Seto, J. (n.d.). *Protein Production*. Biology Oer. <https://openlab.citytech.cuny.edu/bio-oer/protein-production-and-purification/#expression-systems>

[8] Xu, J., Tang, P., Yongky, A., Drew, B., Borys, M. C., Liu, S., & Li, Z. J. (2019). Systematic development of temperature shift strategies for Chinese hamster ovary cells based on short duration cultures and kinetic modeling. mAbs, 11(1), 191–204. https://doi.org/10.1080/19420862.2018.1525262