**Cytovance Biologics CHO Titer Experiment Analysis**

**Data Formatting and Exploration**

Jackson Polk

Summer 2023

DSA 5900-995

4 Credit Hours

Supervisors: Matthew Beattie and Talayeh Razzaghi

Company: Cytovance Biologics

**Preface:**

I have written most code in R as a “scratchpad”. I will go back and re-write in Python once we have verified that these methods are correct.

I have spent a great deal of time working through data formatting issues. Example: pH Setpoint taking a tolerance level of 7.0 +/- 0.2.

**Literature Review:**

Theuse of Machine Learning in tandem with quality control methods to improve titer production in real time has been emphasized (Freier et. al., 2016; Chemmalil et. al. 2021).

Temperature shift during cell cultivation affects both long and short run production; higher temperatures cultivate quickly but die faster, opposite true for lower temperatures (Xu et. al. 2019). Concludes that temperature is an important factor in cultivation, should be focused on during analysis and modeling.

Keeping pH between 6.95 and 7.1 was shown to drastically increase production performance (Ahleboot et. al., 2021).

Carbon Dioxide (CO2) removal is vital to controlling pH in titer production (Ahleboot et. al., 2021).

**Data Formatting Steps:**

1. Standardize Column Names
   1. Convert to lower case
   2. Replace spaces with underscores
   3. Remove special characters
   4. Remove newline characters
   5. Get rid of column names that contain ‘unname’.
2. Standardize Data Formats
   1. Convert character strings to lower case
   2. Remove special characters from character strings
   3. Convert columns to numeric (if applicable).
   4. Convert “Vessel Name” to a categorical variable

**Data Exploration:**

Since extra data preparation has been done, new data exploration figures must be found. Please refer to the Python Notebook “Data Formatting”. Here are the highlights:

CHOKC

|  |  |  |
| --- | --- | --- |
| Vessel Type | Vessel Name | # Observations |
| ambr15 | 1 | 15 |
| ambr15 | 2 | 15 |
| ambr15 | 3 | 15 |
| ambr15 | 4 | 15 |
| ambr15 | 5 | 23 |
| ambr15 | 6 | 23 |
| ambr15 | 7 | 23 |
| ambr15 | 8 | 23 |
| ambr15 | 9 | 27 |
| ambr15 | 10 | 27 |
| ambr15 | 11 | 27 |
| ambr15 | 12 | 27 |
| ambr15 | 13 | 19 |
| ambr15 | 14 | 19 |
| ambr15 | 15 | 19 |
| ambr15 | 16 | 19 |
| ambr15 | 17 | 23 |
| ambr15 | 18 | 23 |
| ambr15 | 19 | 23 |
| ambr15 | 20 | 23 |
| ambr15 | 21 | 27 |
| ambr15 | 22 | 27 |
| ambr15 | 23 | 27 |
| ambr15 | 24 | 27 |

CHOS

|  |  |  |
| --- | --- | --- |
| Vessel Type | Vessel Name | # Observations |
| ambr15 | 1 | 35 |
| ambr15 | 10 | 27 |
| ambr15 | 11 | 27 |
| ambr15 | 12 | 27 |
| ambr15 | 13 | 27 |
| ambr15 | 14 | 27 |
| ambr15 | 15 | 27 |
| ambr15 | 16 | 27 |
| ambr15 | 17 | 27 |
| ambr15 | 18 | 27 |
| ambr15 | 19 | 27 |
| ambr15 | 2 | 35 |
| ambr15 | 20 | 27 |
| ambr15 | 21 | 27 |
| ambr15 | 22 | 27 |
| ambr15 | 23 | 27 |
| ambr15 | 24 | 27 |
| ambr15 | 3 | 35 |
| ambr15 | 4 | 35 |
| ambr15 | 5 | 27 |
| ambr15 | 6 | 27 |
| ambr15 | 7 | 27 |
| ambr15 | 8 | 27 |
| ambr15 | 9 | 27 |
| ambr15 | brx1 | 0 |
| ambr15 | brx2 | 0 |
| epp | 1 | 0 |
| epp | 10 | 0 |
| epp | 11 | 0 |
| epp | 12 | 0 |
| epp | 13 | 0 |
| epp | 14 | 0 |
| epp | 15 | 0 |
| epp | 16 | 0 |
| epp | 17 | 0 |
| epp | 18 | 0 |
| epp | 19 | 0 |
| epp | 2 | 0 |
| epp | 20 | 0 |
| epp | 21 | 0 |
| epp | 22 | 0 |
| epp | 23 | 0 |
| epp | 24 | 0 |
| epp | 3 | 0 |
| epp | 4 | 0 |
| epp | 5 | 0 |
| epp | 6 | 0 |
| epp | 7 | 0 |
| epp | 8 | 0 |
| epp | 9 | 0 |
| epp | brx1 | 30 |
| epp | brx2 | 30 |

Titer by Production Day

CHOKC

A picture containing diagram, line, plot, text

Description automatically generated

CHOS

A picture containing diagram, screenshot, line, design

Description automatically generated

A picture containing diagram, plot, line, text

Description automatically generated

**Algorithm Proposal:**

~~My initial plan is an online, semi-supervised regression model evaluated using root mean-squared error. The use of a regression model assumes that the exact value of titer is needed, rather than a range. The online model is necessitated by the manufacturing process being continuous, it would be helpful to train the model on new data as it comes in. The semi-supervised learning is due to the sparsely-labelled dataset. Finally, the RMSE method will allow proper weighting of outlier predictions.~~

After determining that the input data is largely categorical, regression may not be the best algorithm. However, an “online” model is still preferable, as is a “semi-supervised” one.

Addendum

~~I have been suggested to use reinforcement learning by Dr. Beattie.~~ Dr. Beattie never got back to me on this, and I have decided to ignore the suggestion, respectfully.

**Citations**

Ahleboot Z, Khorshidtalab M, Motahari P, et al. Designing a Strategy for pH Control to Improve CHO Cell Productivity in Bioreactor. Avicenna Journal of Medical Biotechnology. 2021;13(3):123-130. doi:10.18502/ajmb.v13i3.6365

Chemmalil, L, Prabhakar, T, Kuang, J, et al. Online/at-line measurement, analysis and control of product titer and critical product quality attributes (CQAs) during process development. Biotechnology and Bioengineering. 2020; 117: 3757– 3765. <https://doi-org.ezproxy.lib.ou.edu/10.1002/bit.27531>

Freier, L., Hemmerich, J., Schöler, K., Wiechert, W., Oldiges, M. and von Lieres, E. (2016), Framework for Kriging-based iterative experimental analysis and design: Optimization of secretory protein production in Corynebacterium glutamicum. Eng. Life Sci., 16: 538-549. https://doi.org/10.1002/elsc.201500171

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