**Cytovance Biologics CHO Titer Experiment Analysis**

**Project Proposal**

Jackson Polk

Summer 2023

DSA 5900-995

4 Credit Hours

Supervisors: Matthew Beattie and Talayeh Razzaghi

Company: Cytovance Biologics

**Introduction:**

Cytovance Biologics, a biomanufacturing company located in Oklahoma City, is seeking analysis on maximizing the therapeutic protein concentration, or titer from animal cells. Their Chinese Hamster Ovary (CHO) cell line, which produces high quality recombinant and therapeutic proteins, is used in the preclinical to early phase of therapeutic protein development (Cytovance Biologics, n.d., "Mammalian Biologics CDMO Services"). Recombinant proteins are used to instruct a host organism to produce more of a protein that may be lacking. Cytovance is renowned for its expertise in manufacturing these proteins (Cytovance Biologics, n.d., "Complex Molecules") Hence, the goal is to develop an algorithm that analyzes and improves Cytovance’s protein production.

There are two CHO “strains” that Cytovance has experimented on, CHO-S and CHO-KC. Data has been collected at small scale on CHO-S and CHO-KC strain cultivation. Multiple independent variables have been identified and manipulated to put together a dataset that will help train machine learning models. Analysis of experimental data will also allow Cytovance insight on variables that should be thoroughly controlled. Because the manufacturing process is sensitive to measurement, there is limited training data. Proper modeling will provide robust solutions to these issues. A robust model will provide Cytovance with guidance as to relevant independent variables along with a controllable process, driving efficiency in their manufacturing process and ultimately producing higher quality recombinant proteins.

**Objectives:**

The goal is to develop an online, semi-supervised regression model to predict protein titer and optimize the manufacturing process. The use of an online learning model is preferred due to the need to continuously analyze and adjust the model as new data becomes available. Online learning models are well-suited for systems with a stream of information that requires adaptation to changing conditions (Geiron 2019). Since this process is sensitive to measurement, output variables and measurements will be sparse, meaning that this model will need to be semi-supervised. A successful model will accurately predict the protein titer. The final product will allow Cytovance to model their manufacturing process and tweak variables before spending money on potentially ineffective processes.

The collected data will also need to be pre-processed using imputation, outlier detection methods, along with some possible bootstrap sampling for the lack of data. Imputation methods should be carefully considered, as missing values are not easily classified as 0 or any specific number. Outliers will need to be addressed as well. Bootstrap sampling will possibly help create a larger training and test set to train and tune models. The produced model will pose many challenges, mostly by requiring an efficient pipeline for data ingestion. Once data has been properly wrangled and processed, then models can be produced. Most of the work for this project will likely come on the front end, which allows personal knowledge to be developed for online data pipelines.

Altogether, the project allows for real, experimental data to be ingested and utilized in an advanced model. The results will be a robust model that improves the manufacturing process of therapeutic/recombinant proteins.

**Plan:**

Training Data

A full online learning system should be developed that integrates well with the data provided by Cytovance to predict the titer produced by the cultivation process. The training data is experimental and produced by Cytovance. The training data are 2 Microsoft Excel files with 9 sheets for the CHO-S strain and 4 sheets for the CHO-KC strain. There are 1084 total training instances for the CHO-S strain and 800 for the CHO-KC strain. Each dataset contains 14 input variables: Vessel Type, Vessel Name, Production Day, Supplement (in two formats), DO, pH Setpoint, Temperature, Target Cell Seeding Density (cells/mL), Media Type, Feed Type and Feeding Interval. The main target output variable is “Titer by Octet (mg/L)”. There are other output variables that Cytovance would like to know how they are correlated to Titer: Viable Cell Density (cells/mL), Cell Viability (%), Average Cell Diameter (micrometers), pH, Glutamine (mM), Glutamate (mM), Glucose (g/L), Lactate (g/L), Ammonium (mM), Sodium (mM), Potassium (mM), Calcium (mM), Osmolality (mOsm/kg), Bicarbonate (mM), Air Saturation (%) and CO2 Saturation (%).

Data Analysis

To analyze this data, the Cross-Industry Standard Process for Data Mining (CRISP-DM) procedure will be applied. First, as domain knowledge is limited, extensive research should be done. Quick research has revealed that temperature variation during the cultivation process is important to prediction of titer (Xu et. al., 2019). Then, correlation matrices and histograms of input variables will aid with data understanding. The correlation matrices will identify variables that are most important in the modeling procedure, and the histograms will help with understanding data distributions. Missing values and outliers will be identified, and data preparation will then be possible. It is possible that multiple imputation by chained equations (MICE) will be used, as it is preferable to elementary imputation such as mean value imputation. Instead of using a single number, MICE uses the other non-missing variables to build regression models to impute values (Raghunathan et. al. 2000). A simple outlier detection algorithm using standardized data could be used.

Modeling

Rapid modeling and evaluation can then be done with the produced clean data. To evaluate models, Root Mean Square Error (RMSE) will be used, as it weights outlier predictions heavily. Once an adequate model has been identified, a model can be deployed and delivered to Cytovance.

Project deliverables will be presented to Talayeh Razzaghi and conveyed to Cytovance in an effort to fine-tune project findings. Each step will have visualizations, such as a correlation matrix, histograms, and scatter plots. A great visualization that will help Cytovance understand the model is a control chart, which will track the model performance evaluation against time. The technical deliverables for this project are any materials that aid in the analysis of this experiment. Items such as:

1. Correlation Matrix and Data Quality Reports: these will help understand the quality of data that is being modeled, as well as relationships within the data. Any significant assumptions can be graphically depicted in this deliverable, such as linear/non-linear relationships. This deliverable will be in PDF format.

2. Pre-Processing Script: a Python script that will format the data, including imputation and outlier adjustment. This script will be monitored and given to Talayeh Razzaghi.

3. Modeling: various models presented, and evaluations shown in PDF report format. This report will also justify tuning parameters and procedures.

4. Mid-Semester Progress Report: this report will encapsulate every aspect of the project and be presented to the supervisor Matthew Beattie and teaching assistant Elaheh Jafarigol.

5. Final Modeling: the final model and justification documentation will be provided in PDF format.

6. Presentation: a final slide deck and verbal presentation will be given to University of Oklahoma faculty to display knowledge gained.

**Schedule:**

An initial schedule, subject to changes.

May 16th, 2023 Initial Meeting

May 23rd, 2023 Correlation Matrix and Data Quality Reports

May 30th, 2023 Pre-Processing Script

June 6th, 2023 Modeling

June 9th, 2023 Mid-Semester Progress Report

June 13th, 2023 Final Modeling

July 7th, 2023 Project Completion (Presentation Slides/Final Report).

**Citations**

1. Cytovance Biologics. (n.d.). *Mammalian Biologics CDMO Services*. Cytovance. https://cytovance.com/cdmo/mammalian/
2. Cytovance Biologics. (n.d.). *Complex Molecules, Creative Solutions.* Cytovance. <https://cytovance.com/>
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. 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.
5. 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