PROJECT 3

LEO-WYNDOR GLASS PROBLEM

1. INTRODUCTION

The Wyndor glass problem considered for this project falls under the wing of Predictive Stochastic Programming (PSP). It is a combination of both Predictive Analytics and Stochastic Programming. Unlike an SP which has uncertain scenarios associated with it, a PSP works with covariates (the predictor and response variables.) The solving of a PSP model incorporates methodologies from both learning and Optimization. This integration created a combined methodology known as Learning Enabled Optimization (LEO). This methodology is followed for the Wyndor glass problem which requires two problems to be solved - a Budget Allocation and a Production Planning problem.

The problem in discussion considers a company – The Wyndor Glass Co- that produces two different types of high-quality glass doors: Aluminum framed glass doors (y_A) and Wooden framed glass doors (y_B), each with their own requirement of hours. These doors are produced using resources available at three plants 1, 2, and 3.

Table 4 : Production data set

PLANT	PRODUCTION TIME A (Hours/Batch)	PRODUCTION TIME A (Hours/Batch)	TOTAL HOURS	TOTAL HOURS MAGNIFIED (*2)
1	1	0	4	8
2	0	2	12	24
3	3	2	18	36
Profit	$3,000	$5,000		

The problem considers the circumstance that the company can sell up to the number of doors demanded by its customers. These sales could be uncertain and can affect the Profit generated by the company. The sales predictions depend on the 200 advertising slots considered in the problem, through which the company advertises both types of doors. Depending on first stage decisions these 200 slots are to be divided between each outlet- TV (x_1) and Radio (x_2).

Table 5 : Advertising data set

	TV	RADIO	SALES
1	230.1	37.8	22.1
2	44.5	39.3	10.4
…. ..	…. ..	…. ..	…. ..
200	232.1	8.6	13.4

Three statistical methodologies are utilized to solve the PSP Wyndor Glass problem in the discussion: Deterministic Linear Programming (DLP), Sample Average Approximation (SAA), and Stochastic Decomposition (SD).

2. GOAL

The problem in discussion is solved under the circumstance that the company can sell up to the number of doors demanded by its customers. This demand could be affected by advertising which in turn can vary the profit/loss of the company.
The aim is to decide on a budget that is to be spent on each type of outlet- TV, and Radio, to maximize the expected Profit.

3. WORKFLOW

The data obtained is fit to a regression model to determine the Q-Q plot of the errors to determine any outliers. The outliers are then removed, and the linear regression model is fit again. The decisions obtained would be fed to the optimization model to solve the first-stage and second-stage models and get decisions. This model is then validated using Model Validation Sample Average Estimation (MVSAE) against a 95% Confidence Interval to predict the performance in future.

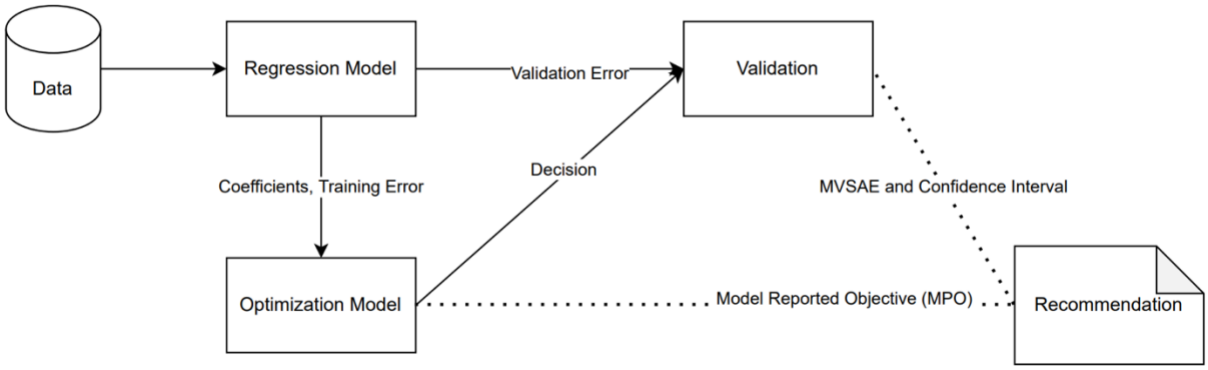

Figure 10 : Workflow

The data is prepared, and the regression model is fit using Pandas, scikit learn library in Python Programming Language in Jupyter Notebook. For the optimization model, Julia programming Language in Jupyter Notebook is used with CPLEX Solver.

4. METHODOLOGY

The problem in discussion utilizes two stage solving hence two models. The first model would be the budget allocation model. The variables $x_1$and x_2 are defined as expenditures for TV and Radio outlets respectively. The total budget allocated is given by b = 200 in thousands of dollars. The first stage model is given by

$$Max \ -0.1x_1 - 0.5x_2 + \mathbb{E}[Profit(\omega)]$$

$$s.t \ x_1 + x_2 \leq 200$$

$$x_1 - 0.5x_2 \geq 0$$

$$L_1 \leq x_1 \leq U_1, L_2 \leq x_2 \leq U_2$$

Equation 9 : First stage model

The first constraint satisfies the requirement that total expenditure on the outlets must be less than the allocated budget. The second constraints make sure that atleast some amount is spent on both the outlets. The upper and lower limits on the variables is to include only those predictions that fall within the range of expenditures for the data in the Advertising data set.

The ω denotes the sales and is given by $\omega = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. The coefficients ($\{\beta_j\}$) are treated as random variables because there are estimation errors associated with them. This ω is defined in the second stage model as follows

$$Profit(\omega_i) = Max \ 3y_A + 5y_B$$

$$s.t. \quad y_A \leq 8$$

$$2y_B \leq 24$$

$$3y_A + 5y_B \leq 36$$

$$y_A + y_B \leq \beta_0 + \beta_0 x_1 + \beta_0 x_2 + \varepsilon_{ti}, \quad \forall_i$$

$$y_A,\ y_B \geq 0$$

Equation 10: Second stage model

4.1 Linear Regression

To perform the Linear regression, the data set is randomly split into training and validation sets each with 100 data points. A linear regression is then fit to the training data set using Sales as the response variable with TV, Radio as predictor variables. The values obtained are recorded as the intercept (β_0), coefficient of the TV outlet (β_1), and the coefficient for the Radio outlet (β_2).

```
#beta1,beta2
cdf = pd.DataFrame(lm.coef_, Xtrain.columns, columns=['Coefficients'])
print(cdf)
beta1,beta2 = lm.coef_

#beta0
beta0 = lm.intercept_
```

```
       Coefficients
TV         0.045908
Radio      0.201280
```

```
beta0,beta1,beta2
```

```
(2.60361291935199, 0.04590841383514319, 0.20127960643138415)
```

Figure 11 : Linear regression model

For the next step, we calculated the training and Validation errors from respective datasets for each data point $(x_{1i}, x_{2i}, \omega_i)$ using the following equations

$$\varepsilon_{ti} = \omega_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}$$

Equation 11 : training error calculation:

$$\varepsilon_{vi} = \omega_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}$$

Equation 12 : Validation error calculation

The training and validation errors and ω are obtained as follows

Figure 12 : ε_{ti}

```
43    -0.232381
72     1.704720
160    0.515659
14     0.777709
157   -0.201023
          ...
164    0.829736
180    0.365570
91     1.237176
137   -2.286798
136   -2.640601
Length: 100, dtype: float64,
```

Figure 13 : ε_{vi}

```
59     2.242164
186    2.129535
78     0.392982
116   -0.517401
159   -3.508906
          ...
77     2.044584
139   -0.462306
171    0.724254
126    1.323257
200   -1.366951
Length: 100, dtype: float64,
```

Figure 14 : ω

```
59     18.242164
186    18.129535
78     16.392982
116    15.482598
159    12.491094
          ...
77     18.044584
139    15.537694
171    16.724254
126    17.323257
200    14.633049
Length: 100, dtype: float64
```

An F-test is performed using the above-calculated error samples.

```
import scipy.stats

# Create data
group1 = eti
group2 = evi

# converting the list to array
x = np.array(group1)
y = np.array(group2)

# calculate variance of each group
print(np.var(group1), np.var(group2))

def f_test(group1, group2):
    f = np.var(group1, ddof=1)/np.var(group2, ddof=1)
    nun = x.size-1
    dun = y.size-1
    p_value = 1-scipy.stats.f.cdf(f, nun, dun)
    return f, p_value

# perform F-test
f_test(x, y)
```

Figure 15 : F-test

```
2.516081195233829 3.1309930242818633
(0.803604854983964, 0.8608042387998502)
```

Figure 16 : p-value

The p-value is 3.13 approximately which is greater than 0.05 which indicates that the null hypothesis can be rejected. A Q-Q Plot is then generated to check the presence of any outliers for both training and validation errors.

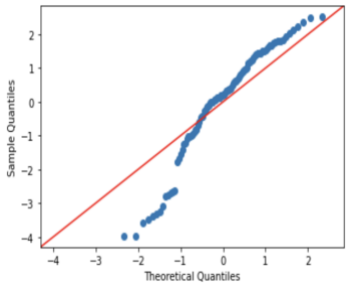

Figure 17: Q-Q plot for training errors

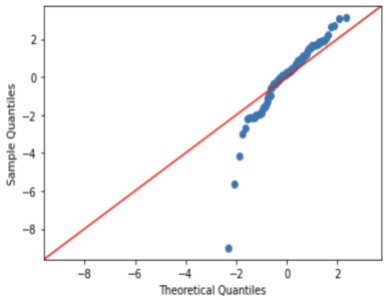

Figure 18: Q-Q plot for validation errors

4.2 Deterministic Linear Programming

The first statistical technique used is the Deterministic Linear Programming. The model assumes the training error due to regression as zero hence the sales predictions for this model are given by

$$m(x, \varepsilon) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

To solve the problem using the deterministic approach the two-stage optimization problem is converted into an equivalent All-in-one model as follows

$$\begin{aligned}
Max \ & -0.1x_1 - 0.5x_2 + 3y_A + 5y_B \\
s.t \ & x_1 + x_2 \leq 200 \\
& x_1 - 0.5x_2 \geq 0 \\
& y_A \leq 8 \\
& 2y_B. \leq 24 \\
& 3y_A + 5y_B \leq 36 \\
& \beta_1 x_1 + \beta_2 x_2 + y_A + y_B \leq \beta_0 \\
& y_A, y_B \geq 0 \\
& L_1 \leq x_1 \leq U_1, L_2 \leq x_2 \leq U_2
\end{aligned}$$

Equation 13: Integrated model for deterministic optimization

In the above model, the random variable is replaced by its expectation as we are considering a deterministic approach. The decision variables are the same as in the two-stage model and the constraints are only slightly changed. The value of ω is replaced and re-arranged in the sixth constraint.

4.3 Slp With Sample Average Approximation And Stochastic Decomposition Model

The second and third statistical techniques used are Sample Average Approximation and Stochastic Decomposition respectively. These techniques consider the errors obtained by the regression. These errors are considered as random variables. The sales predictions equation is given as follows

$$m(x, \varepsilon) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_{ti}$$

Equation 14

The error term represents the training errors generated by fitting the linear regression model to the training data set. The model used for the SAA is as follows

$$Max \; -0.1x_1 - 0.5x_2 + \frac{1}{N}\sum_{i=1}^{N}(3y_{Ai} + 5y_{Bi})$$

$$\begin{aligned}
s.t \; x_1 + x_2 &\leq 200 \\
x_1 - 0.5x_2 &\geq 0 \\
y_{Ai} &\leq 8 \qquad i = 1, \dots\dots\dots, N \\
2y_{Bi}. &\leq 24 \qquad i = 1, \dots\dots\dots, N \\
3y_{Ai} + 5y_{Bi} &\leq 36 \qquad i = 1, \dots\dots\dots, N \\
-\beta_1 x_1 - \beta_2 x_2 + y_A + y_B &\leq \beta_0 + \varepsilon_{ti} \qquad i = 1, \dots\dots\dots, N
\end{aligned}$$

$$L_1 \leq x_1 \leq U_1, L_2 \leq x_2 \leq U_2, y_{Ai}, y_{Bi} \geq 0$$

Equation 15

In the above model, the sixth constraint incorporates the errors generated from the training set.

5. VALIDATION

The outcomes from the above-mentioned three techniques are validated to understand each model's performance to compare. The validation is done using the Model Validation Sample Average Estimation (MVSAE). In this, we use the validation data set to solve the problem. A 95% confidence interval is generated to check if the training results fall inside the interval.

The DF, SAA, and SD are solved using equations 5 and 7 respectively. The first staged decisions are then fixed and the second stage decisions are calculated. The Mean Objective value (MPO),

the standard deviation, and the 95% confidence interval is calculated. The results obtained from the training set are compared with the confidence intervals.

6. RESULTS

The results generated from the three models are mentioned in the table below.

MODEL	X1 (TV)	X2 (Radio)	MPO	MVSAE
Deterministic	172.8733	27.1266	$[41.1493]	$[38.6634,39.7148]
SLP with SAA	181.3617	18.6382	$[39.9318]	$[39.6493,40.2144]
SLP with SD	181.3617	18.6382	$[39.9509]	$[39.6682,40.2335]

Table 6: Results

From the results obtained, we recommend SLP with SAA and SLP with SD as better performing models. This is because the MPO obtained is within the validation MVSAE interval.

7. FUTURE WORK

As future work, our team would like to use different splits of training and validation sets and fit the regression to the training data. We would also like to study other validation approaches to compare with the MVSAE used in this discussion.

8. REFERENCES

[1] Suvrajeet Sen, Jiajun Xu, Yihang Zhang, "ISE 533 Notes : Integrative Analytics with Cross-Sectional Data", ISE533 Blackboard, 2022.
[2] Y.T. Herer, M. Tzur, and E. Yucesan, "The multilocation transshipment problem", IIE Transactions, Apr. 2006.
[3] Group 7 ppt, "WYNDOR GLASS PROBLEM", presented in class.