

#SASGF

The logo for the Virtual SAS Global Forum 2021. The word "VIRTUAL" is written in a large, bold, white, sans-serif font. Each letter of "VIRTUAL" contains a colorful, abstract pattern of diagonal stripes in shades of blue, red, green, and purple. Below "VIRTUAL" is the text "SAS® GLOBAL FORUM 2021" in a smaller, white, sans-serif font. The entire logo is centered on a dark blue background.

VIRTUAL
SAS® GLOBAL FORUM 2021

Opening the door to next generation data mining & machine learning – Python & SAS together at last!

George S. Habek, M.S. – CT Global Solutions, Inc.



George has been using SAS for over thirty years and has an extensive background in Retail, Manufacturing, Hospitality & Entertainment, Pharmaceutical, Automotive, Financial, Restaurant/Food, Agriculture & Construction, and much more. His recent accomplishments include working as a senior analytical trainer at SAS (2006-2017) and teaching business analytics at North Carolina State University.

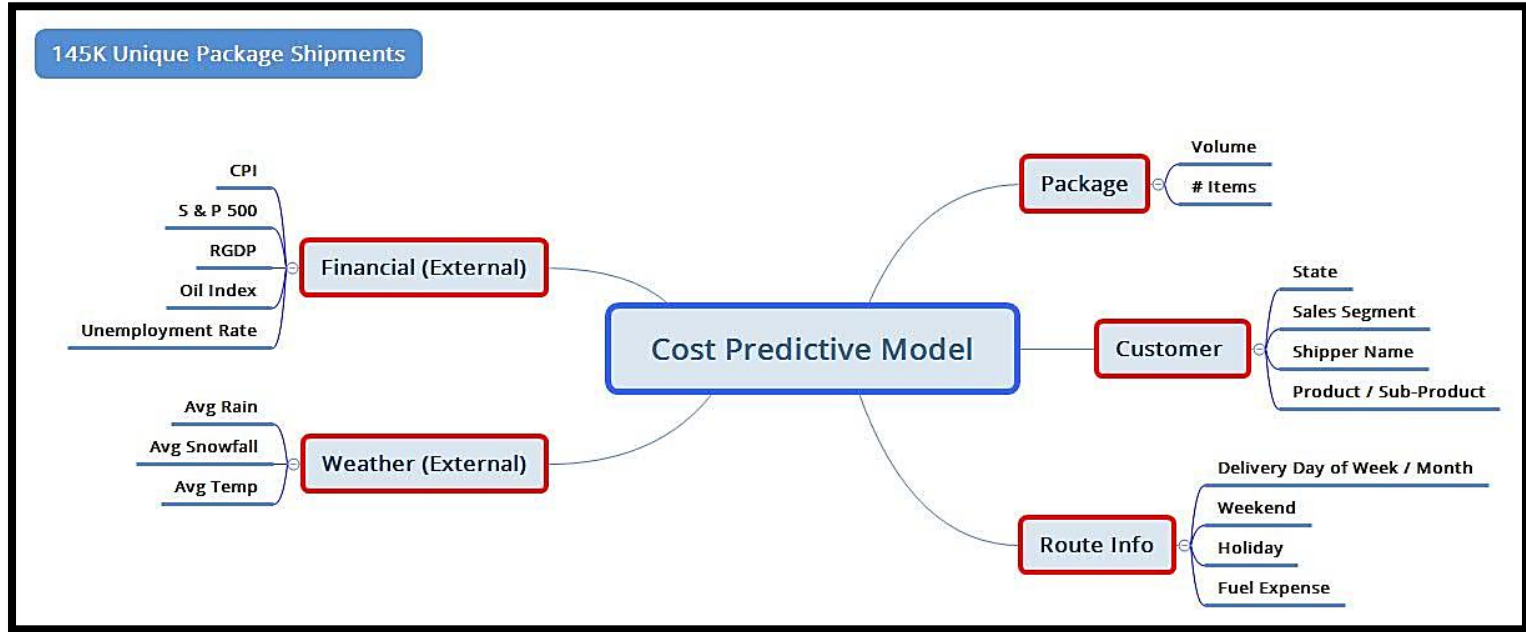
Introduction

Business Problem

- Transportation logistics industry.
- The ultimate business question – What is the best subset of factors that drive shipping cost? What is the optimal model to use to score a new prospect file with the predicted average cost?
- A variety of data elements such as internal factors of package size, delivery day of the week, # delivery days, etc....; external factors of weather and financial effects will also be used.

The Data

Available Data for the Business Problem



Open-Source Connection

Install & Configure Python 2.7x or 3.4+

- Python should be installed in the Compute server of SAS Viya environment.
- Executable file must be available in system path. Connect to python from SAS VDMML, set the environment variable in `/opt/sas/viya/config/etc/sysconfig/compsrv/default/sas-compsrv` file and add the line `export PATH=path_to_your_python_bin_directory:${PATH}`.
- Give `sudo` privileges to the python packages.

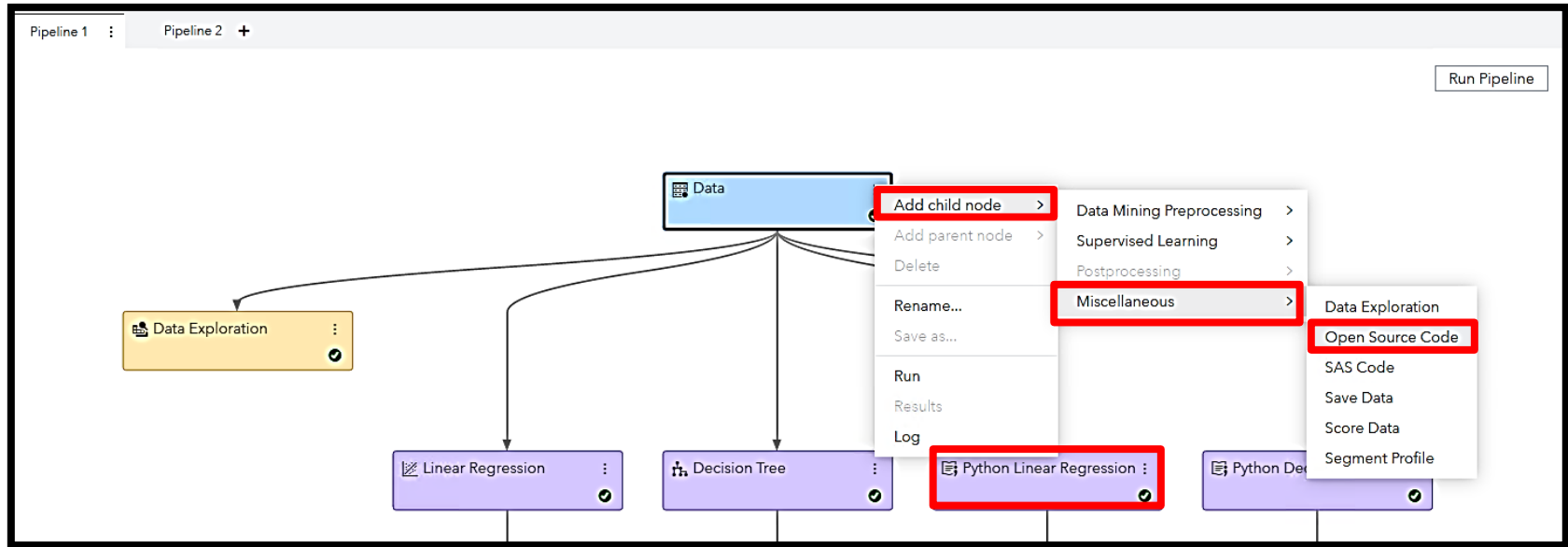
Python Models

Develop Python Models via Jupyter Notebook

- Four machine learning models were developed within Python.
- Sklearn library was utilized for the models.
- Linear Regression using *ElasticNet* algorithm.
- Neural Network
- Decision Tree
- Random Forest

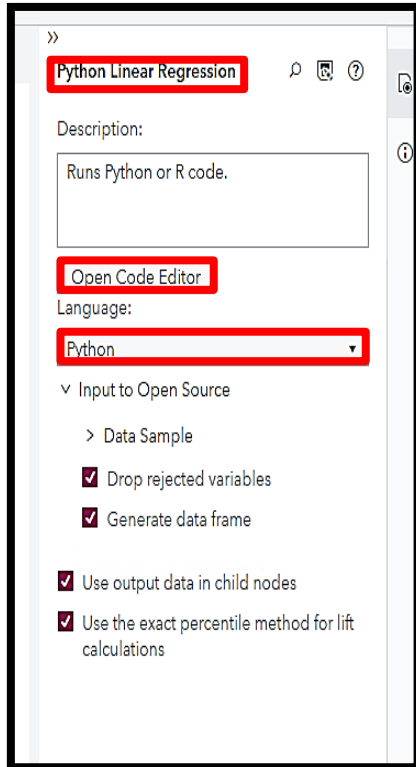
Python & SAS Models

Incorporate Python Models within SAS VDMML Pipeline



Python & SAS Models

Incorporate Python Models within SAS VDMML Pipeline



The screenshot shows the configuration for a Python Linear Regression model in SAS VDMML. The title bar reads "Python Linear Regression". The description is "Runs Python or R code." The "Open Code Editor" button is highlighted. The language is set to "Python". Under "Input to Open Source", "Data Sample" is selected. Checkmarks are present for "Drop rejected variables", "Generate data frame", "Use output data in child nodes", and "Use the exact percentile method for lift calculations".

Python Linear Regression

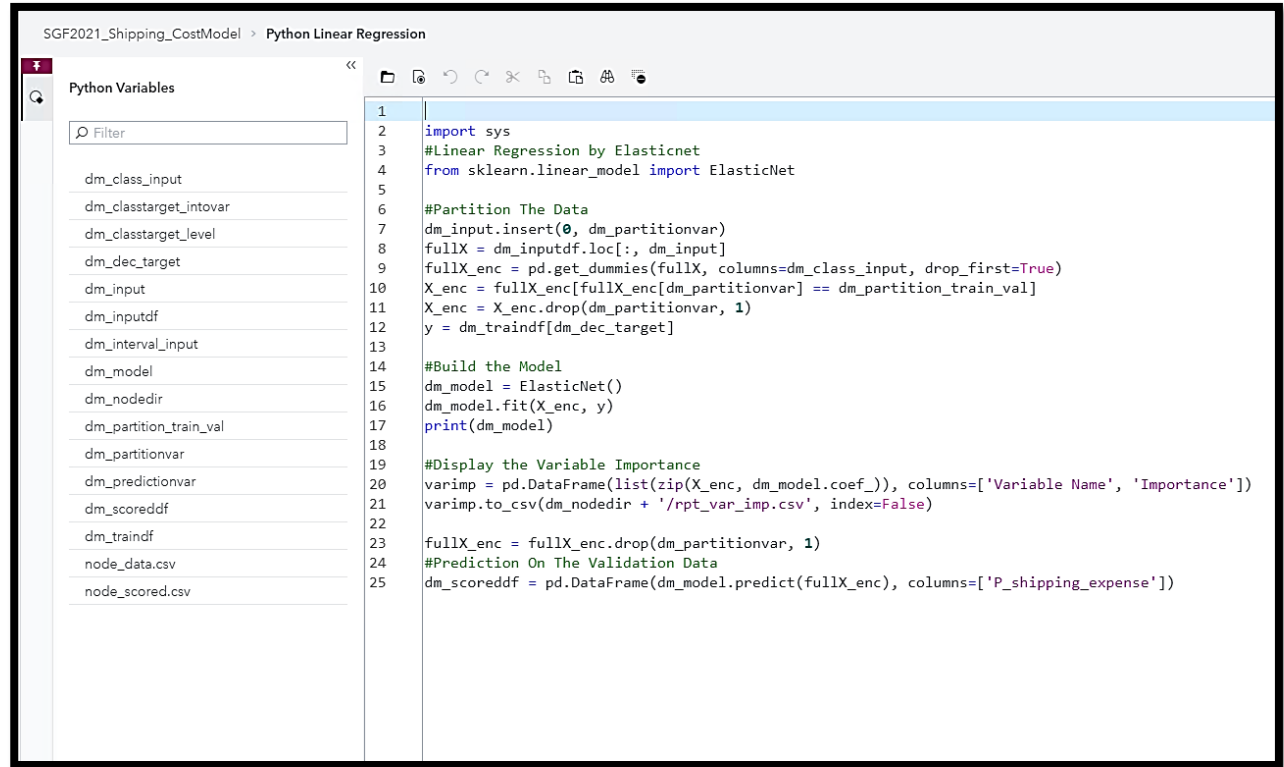
Description:
Runs Python or R code.

Open Code Editor

Language:
Python

Input to Open Source

- Data Sample
- Drop rejected variables
- Generate data frame
- Use output data in child nodes
- Use the exact percentile method for lift calculations



The screenshot shows the code editor for the Python Linear Regression model. The title bar reads "SGF2021_Shipping_CostModel > Python Linear Regression". The code is as follows:

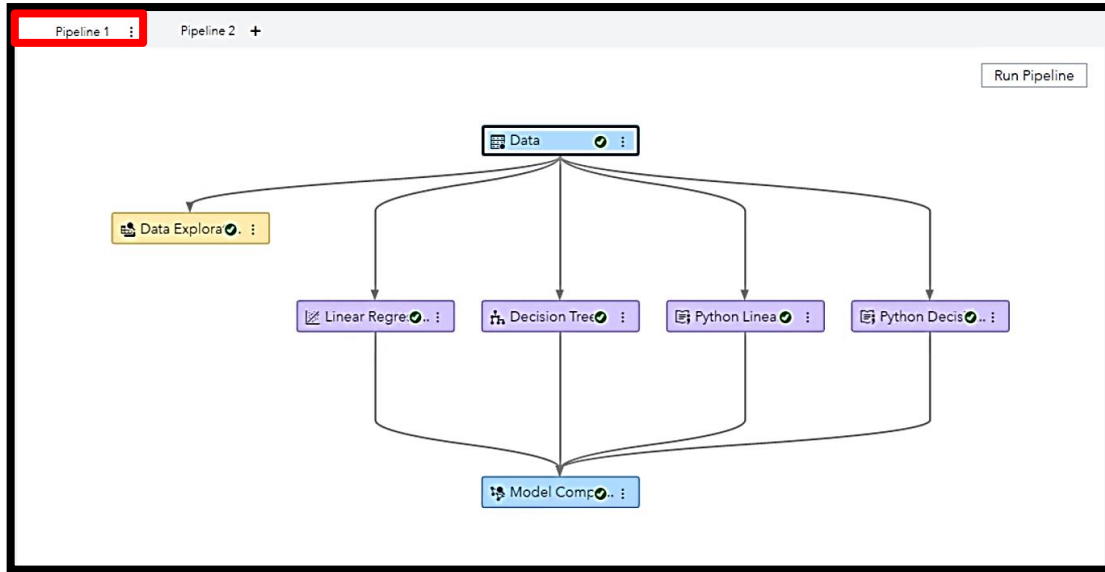
```
1
2 import sys
3 #Linear Regression by Elasticnet
4 from sklearn.linear_model import ElasticNet
5
6 #Partition The Data
7 dm_input.insert(0, dm_partitionvar)
8 fullX = dm_inputdf.loc[:, dm_input]
9 fullX_enc = pd.get_dummies(fullX, columns=dm_class_input, drop_first=True)
10 X_enc = fullX_enc[fullX_enc[dm_partitionvar] == dm_partition_train_val]
11 X_enc = X_enc.drop(dm_partitionvar, 1)
12 y = dm_traindf[dm_dec_target]
13
14 #Build the Model
15 dm_model = ElasticNet()
16 dm_model.fit(X_enc, y)
17 print(dm_model)
18
19 #Display the Variable Importance
20 varimp = pd.DataFrame(list(zip(X_enc, dm_model.coef_)), columns=['Variable Name', 'Importance'])
21 varimp.to_csv(dm_nodedir + '/rpt_var_imp.csv', index=False)
22
23 fullX_enc = fullX_enc.drop(dm_partitionvar, 1)
24 #Prediction On The Validation Data
25 dm_scoreddf = pd.DataFrame(dm_model.predict(fullX_enc), columns=['P_shipping_expense'])
```

Python Variables

- dm_class_input
- dm_classtarget_intovar
- dm_classtarget_level
- dm_dec_target
- dm_input
- dm_inputdf
- dm_interval_input
- dm_model
- dm_nodedir
- dm_partition_train_val
- dm_partitionvar
- dm_predictionvar
- dm_scoreddf
- dm_traindf
- node_data.csv
- node_scoreddf.csv

Python & SAS Models

Incorporate Python Models within SAS VDMML Pipeline

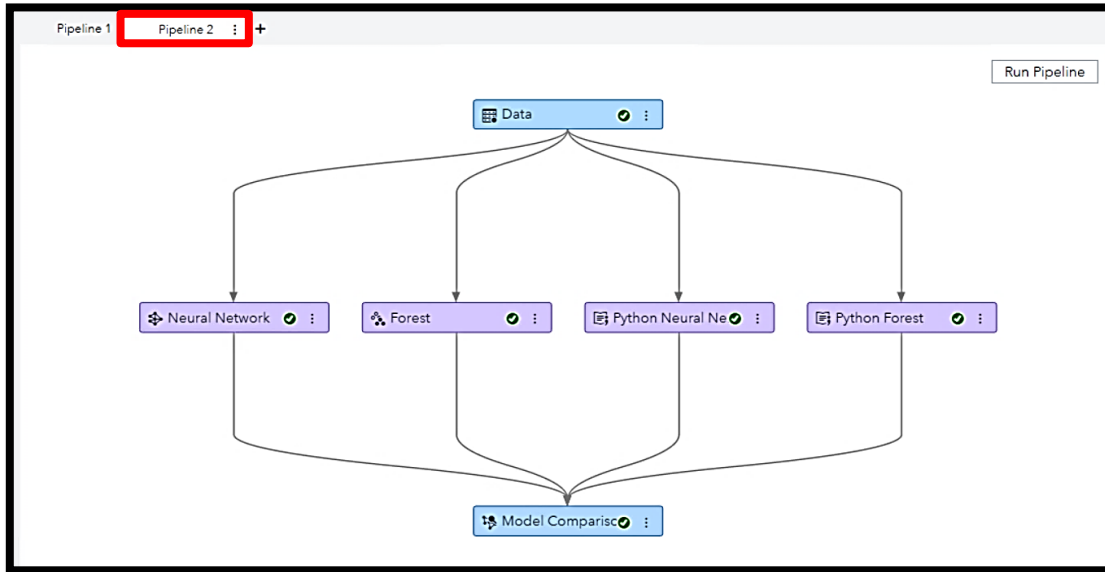


SAS VDMML Linear Regression is the champion for Pipeline 1

Model Comparison				
Champion	Name	Algorithm Name	Average Squared Error	Root Average Squared Error
	Linear Regression	Linear Regression	333.8898	18.2727
	Python Linear Regression	Open Source Code	412.7342	20.3159
	Python Decision Tree	Open Source Code	715.2395	26.7440
	Decision Tree	Decision Tree	835.3924	28.9032

Python & SAS Models

Incorporate Python Models within SAS VDMML Pipeline



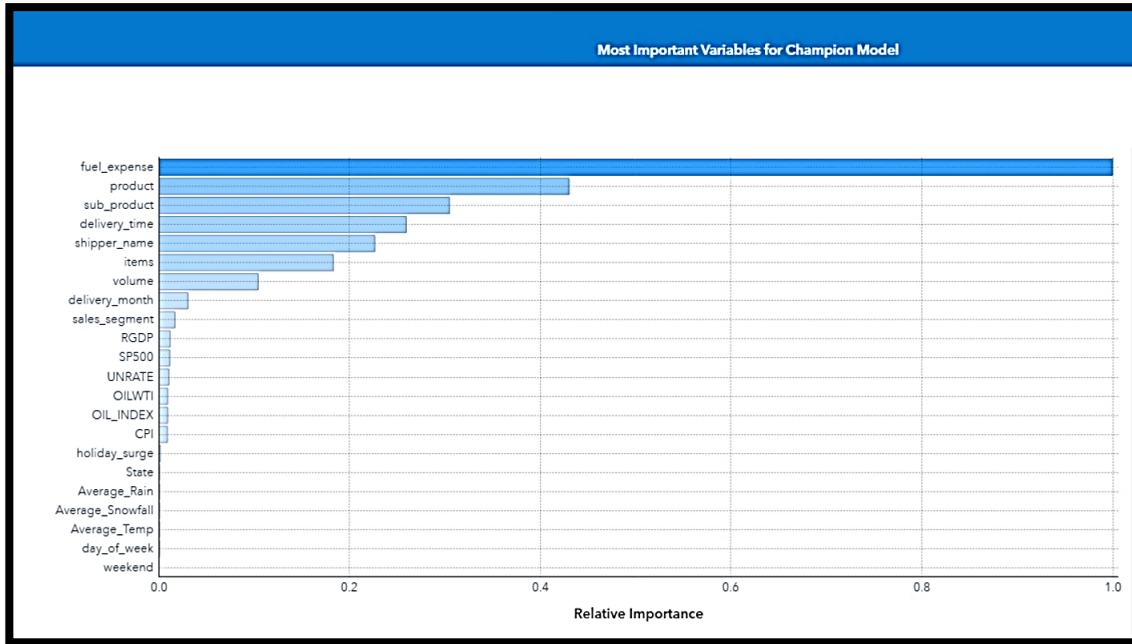
Python Random Forest is the champion for Pipeline 2

Champion	Name	Algorithm Name	Average Squared Error	Root Average Squared Error
	Python Forest	Open Source Code	297.8433	17.2581
	Python Neural Network	Open Source Code	580.6186	24.0960
	Forest	Forest	928.6298	30.4734
	Neural Network	Neural Network	8,309.5219	91.1566

Model Comparison

Compete Python & SAS VDMML Models

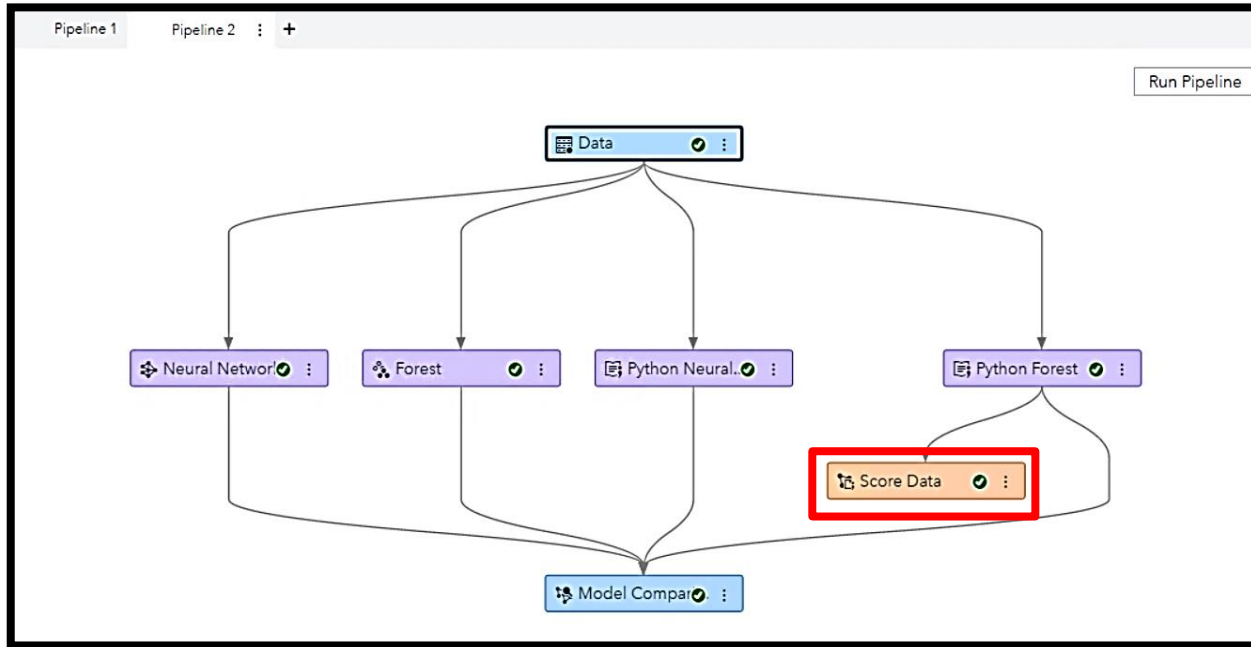
Champion	Name	Algorithm Name	Pipeline Name	Average Squared Error	Sum of Frequency
<input checked="" type="checkbox"/>	Python Forest	Open Source Code	Pipeline 2	297.843	43,903
<input type="checkbox"/>	Linear Regression	Linear Regression	Pipeline 1	333.890	43,903



**Python Random Forest
Overall Champion –
Significant Factors Driving
Shipping Expense**

Model Deployment

Score Prospect File Using Champion Model



**Score New Prospect File
Based on Champion Model**



**Predict Avg Shipping
Expense for Each Package
Shipment**

Model Deployment

Score Prospect File Using Champion Model

State	Sales Segment	Shipper Name	Product	Sub Product	Items	Volume	Fuel Expense ▲	Delivery Time (Days)	Shipping Expense	Predicted Shipping Expense
California	Small/Medium Business	Nautlius SportsWear Inc	3 Day Select	Residential Pick Up/D...	11	3.76	\$2.14	3	\$22.85	\$22.95
Georgia	Small/Medium Business	AllSeasons Outdoor Clothing	3 Day Select	Do Not Stack	40	1.85	\$2.14	3	\$18.22	\$22.27
Indiana	Amazon	3Top Sports	Surepost	Do Not Stack	9	19.15	\$2.14	0	\$24.86	\$24.78
Washington	Small/Medium Business	Mayday Inc	Ground Residential	Do Not Stack	10	4.52	\$2.14	0	\$22.47	\$23.44
Wisconsin	Small/Medium Business	Luna sastreria S.A.	3 Day Select	None	9	12.27	\$2.14	3	\$22.10	\$23.04
Michigan	Enterprise Accounts	Top Sports	Ground Commercial	None	24	37.32	\$2.14	0	\$22.94	\$23.36
South Carolina	Enterprise Accounts	AllSeasons Outdoor Clothing	Hundredweight (CWT)	Notifications	33	135.35	\$2.14	0	\$25.59	\$31.44
Michigan	Small/Medium Business	A Team Sports	Ground Commercial	Lift Gate	18	42.69	\$2.14	0	\$26.40	\$24.33
California	Small/Medium Business	SD Sporting Goods Inc	Surepost	Weekend Delivery	18	3.27	\$2.14	3	\$23.25	\$22.74
Iowa	Enterprise Accounts	Fuller Trading Co.	3 Day Select	None	8	16.45	\$2.14	3	\$28.99	\$28.50
Indiana	Small/Medium Business	Luna sastreria S.A.	Ground Residential	Notifications	7	0.50	\$2.15	0	\$24.31	\$26.42
Arizona	Enterprise Accounts	AllSeasons Outdoor Clothing	Ground Commercial	None	10	29.77	\$2.15	0	\$35.11	\$24.77
Washington	Enterprise Accounts	Eclipse Inc	Ground Commercial	None	14	5.30	\$2.15	0	\$24.98	\$25.70
New Jersey	Enterprise Accounts	Eclipse Inc	Ground Commercial	Weekend Delivery	8	5.41	\$2.15	0	\$24.69	\$25.88
New Mexico	Small/Medium Business	Miller Trading Inc	3 Day Select	Do Not Stack	16	75.79	\$2.15	3	\$26.49	\$26.27
Ohio	Small/Medium Business	Mike Schaeffer Inc	Hundredweight (CWT)	Lift Gate	27	216.89	\$2.15	0	\$9.53	\$23.93
Washington	Enterprise Accounts	Fuller Trading Co.	Ground Commercial	Do Not Stack	8	7.51	\$2.15	0	\$29.15	\$25.98
Florida	Small/Medium Business	SD Sporting Goods Inc	Ground Commercial	Do Not Stack	18	60.13	\$2.15	6	\$25.37	\$24.72
Ohio	Small/Medium Business	Nautlius SportsWear Inc	3 Day Select	Lift Gate	10	3.60	\$2.16	3	\$20.53	\$22.64
Texas	Small/Medium Business	Luna sastreria S.A.	Ground Commercial	Limited Access	18	75.94	\$2.16	0	\$22.47	\$23.99
New York	Small/Medium Business	British Sports Ltd	Ground Commercial	Limited Access	22	65.21	\$2.16	0	\$28.07	\$27.05
North Carolina	Small/Medium Business	Green Lime Sports Inc	Ground Residential	None	10	8.69	\$2.16	0	\$22.79	\$22.36
Pennsylvania	Amazon	3Top Sports	3 Day Select	Lift Gate	10	3.92	\$2.16	3	\$22.05	\$23.66
Massachusetts	Small/Medium Business	Dolphin Sportswear Inc	3 Day Select	Do Not Stack	18	95.54	\$2.16	3	\$21.85	\$24.48
Kansas	Amazon	3Top Sports	Ground Residential	Limited Access	22	48.49	\$2.16	0	\$27.26	\$27.27
Utah	Small/Medium Business	Nautlius SportsWear Inc	Surepost	Do Not Stack	10	11.53	\$2.16	5	\$22.30	\$25.03
Maryland	Small/Medium Business	CrystalClear Optics Inc	Surepost	Weekend Delivery	9	16.12	\$2.16	1	\$22.37	\$23.06
Illinois	Small/Medium Business	Scandinavian Clothing A/S	Surepost	Weekend Delivery	9	15.26	\$2.16	1	\$23.21	\$26.30
Georgia	Small/Medium Business	Van Dammeren International	Ground Commercial	None	8	18.70	\$2.16	0	\$27.94	\$23.53

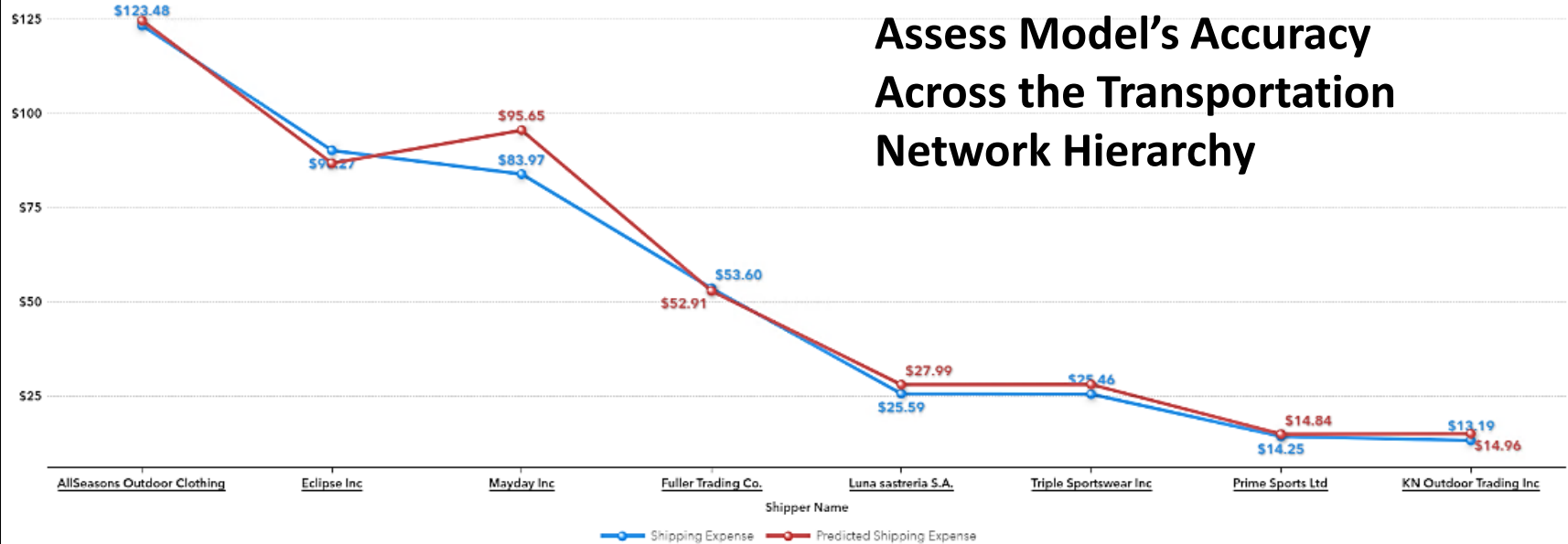
Model Deployment

Score Prospect File Using Champion Model

Actual vs Predicted Shipping Expense - Champion Model

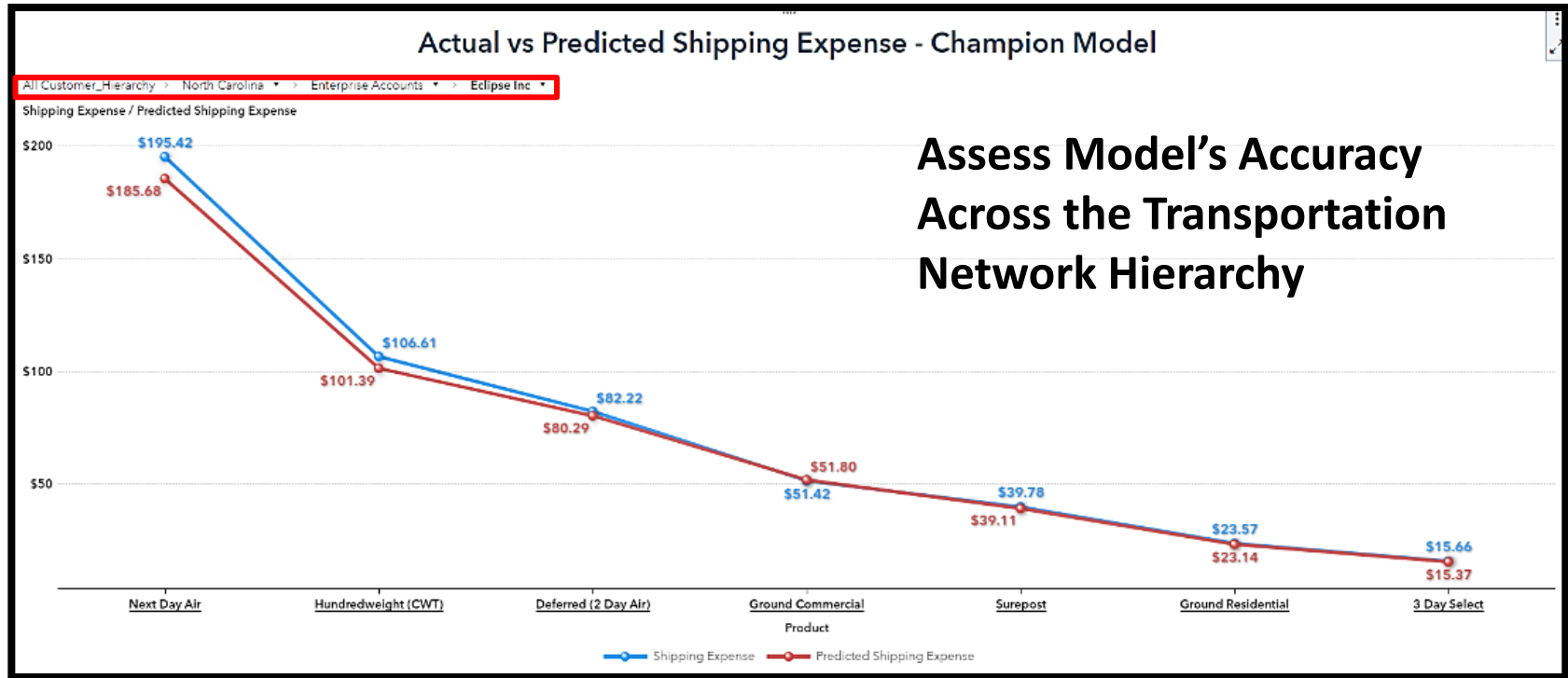
All Customer_Hierarchy > North Carolina > Enterprise Accounts

Shipping Expense / Predicted Shipping Expense



Model Deployment

Score Prospect File Using Champion Model



Conclusion

Summary & Recommendations

- Shipping organization has a critical need to determine causal factors for cost of package shipments across the transportation network.
- Understanding these factors helps the company establish pricing for their customers to become more profitable.
- Integrating open-source models within a next generation data mining & machine learning environment, allowed for a very comprehensive analysis to address the critical business need.
- Capture other data relevant to package shipments.
- Inject the shipping expense causal factors from the champion ML & AI model into a time series forecast.

#SASGF

SAS[®] GLOBAL FORUM 2021

sasglobalforum.com