Adventure Works Cycles:

An Analytics-Based Proposal

Josh Cubero

MSBA 305

Golden Gate University

Abstract

The business landscape today is the result of years of evolution in business strategy that has been influenced by the emergence of modern technologies in data and analytics. These technologies have enabled tech juggernauts such as Amazon, Netflix, Google, and Facebook to achieve great heights and figuratively write the book on how to effectively employ data and analytics. While these companies have mastered the use of data and analytics, many companies are still growing and learning how to employ such strategies. This paper will analyze a small manufacturing company by applying descriptive, predictive, and prescriptive analytics. The outcome will be a machine learning model that can be used to prescribe a number of units to produce based on the desired result.

Literature Review

The literature review in this study was conducted by a team of researchers at the College of Wooster and is entitled "*Sales Forecasting Using Regression and Artificial Neural Networks."* The goal of the study was to combine linear regression methodologies with artificial neural networks (ANN) to predict sales at a manufacturing firm. The study utilized historical data to create the linear model, and ANN to smooth seasonal and non-seasonal data (Nguyen, et. al., 2015). The time series in the study were predictive that spanned from one calendar quarter to 20 calendar quarters. The data set in the study differed from the AWC dataset in that the former incorporates micro and macro-economic factors external to the company, however the principal methods are indeed similar.

According to Nguyen et. al., the ability to forecast sales is an integral part of any successful firm's strategy. Companies continue to find new methods of forecasting sales to gain a

competitive advantage in industry (Nguyen et., al 2015). The research team contends that effective forecasting models enable manufacturers to maintain a proper amount of manufacturing material to satisfy customer orders. This principle is founded in Taiichi Ohno's Just In-Time manufacturing philosophy developed in the 1970s. The philosophy of JIT can be summed up to say that the effective manufacturer minimizes waste by possessing the right materials at the right time to satisfy customer needs (IFM, 2018). The study in this literature review supports the JIT philosophy by creating and evaluating a methodology that enabled the team to predict product sales for a real-world company. The company's name was masked to prevent leakage of company trade secrets. Additionally, the research team did not disclose units of measure, as these were not germane to the model's ability to forecast results.

The model in this literature review had immaterial differences from the AWC model in that their model used micro and macro-economic data. The nature of these data differed in seasonality from the historical sales data. With this in mind, the researchers de-seasonalized the sales history data for the study. This smoothed the historical data, so it matched the patterns of the economic data. The research team used a method called logically weighted scatterplot smoothing to smooth the economic data and is a quadratic regression technique. The pre-smoothed model exhibited significant seasonality, which reduced through smoothing.

The next phase of the study involved choosing economic factors for the model. The researchers stated that there were approximately ninety economics factors which raised a concern that multicollinearity might exist in the data model. The ability of machine learning models to aid in understanding predictions is of most importance in any regression model (Section, 2021). However, regression models can be impacted by variables in the study and become unreliable. A common issue in regression models, multicollinearity. Multicollinearity is a condition that affects

the independent variable/s in a regression model. In a perfect model, independent variables will have little to no correlation. Multicollinearity exists when independent variables in a regression model exhibit correlation (Section, 2015). Multicollinearity is a problem because the condition inflates the statistical significance of the independent variables, which overstates the extent to which the independent variables influence the dependent variable, and not chance alone (Section, 2015). The researchers in this study used statistical techniques that included R-squared, student's T, correlation matrix, and variance inflation factor.

The regression model was used to determine which economic factors had the greatest impact on the predictive capabilities of the model. The regression model enables researchers to estimate the extent to which independent variables affect the outcome of the dependent variable (Nyguen, 2015). The dependent variable in this review's study is sales, and the regression analysis looks at the strength of the relationship between the dependent and independent variables.

The next important analysis in the literature review is the R-Squared value. R-Squared in a regression model is a statistical measure that assesses the amount of variance in the dependent variable that can be attributed to the independent variables (CFI, 2022). R-Squared is commonly referred to as the goodness of fit. R-Squared translates to the percentage of the dependent variable's variance that is caused by the independent variables. For example, if a model's R-Squared is 0.8912, we might say that 89% of the variation is caused by the independent variables. R-Squared is important because a low R-Squared value increases the likelihood that variance in the dependent variable is caused by chance and not by the independent variables (Duke, 2022).

The researchers in this study created a correlation matrix to identify variables with correlation. The correlation matrix runs regression on input variables and presents them in a matrix that enables the researchers to observe their correlation. With correlated dependent variables identified, the team conducted a variance inflation factor exercise to determine the extent of variance inflation in the identified independent variables. The team started the exercise with ninety economic variables that could explain variation in the sales variable. The ninety economic variables were reduced to fifty-six variables whose correlation to sales was greater than 0.50. The team further reduced the independent variables to eighteen via correlation matrix and variance inflation factor analysis. Next the team utilized industry knowledge and judgement to reduce the independent variables to ten.

With the dataset finalized, the team moved to creating the machine learning model. This process involved three stages: training, testing, and validation. The process involves splitting the dataset into to variables, typically X and y, where X is a scalar value or and two-dimensional array of values, and y is the dependent variable (PSU, 2015). The X and y variables are then input as arguments to a machine learning algorithm that further splits the variables into training and test variables. The training variables were used to train the model, while the test variables were used to assess the model for accuracy.

There were three primary outcomes for this study. The first was that the researchers were able to prove that they can remove seasonality from sales data, which essentially smoothed the data. Next, the researchers were able to use statistical methods to conduct dimension reduction from ninety independent variables to ten independent variables. Lastly, after smoothing and dimension reduction, the researchers then executed multiple tests to derive percentage error with results ranging from 5 to 10% error. The researchers also stated that this study was indeed

repeatable in other industries, albeit with different variables (Nyguen et. al, 2015). Furthermore, the researchers concluded that additional research aimed at automating the forecasting methodologies in this study.

The principles used in the literature review will also be applied in the AWC analysis. The plan for this study will include dimension reduction through statistical analysis. Next, the data will be smoothed through grouping and sum functions. Finally, predictive, and prescriptive models will be with a linear regression-based machine learning model.

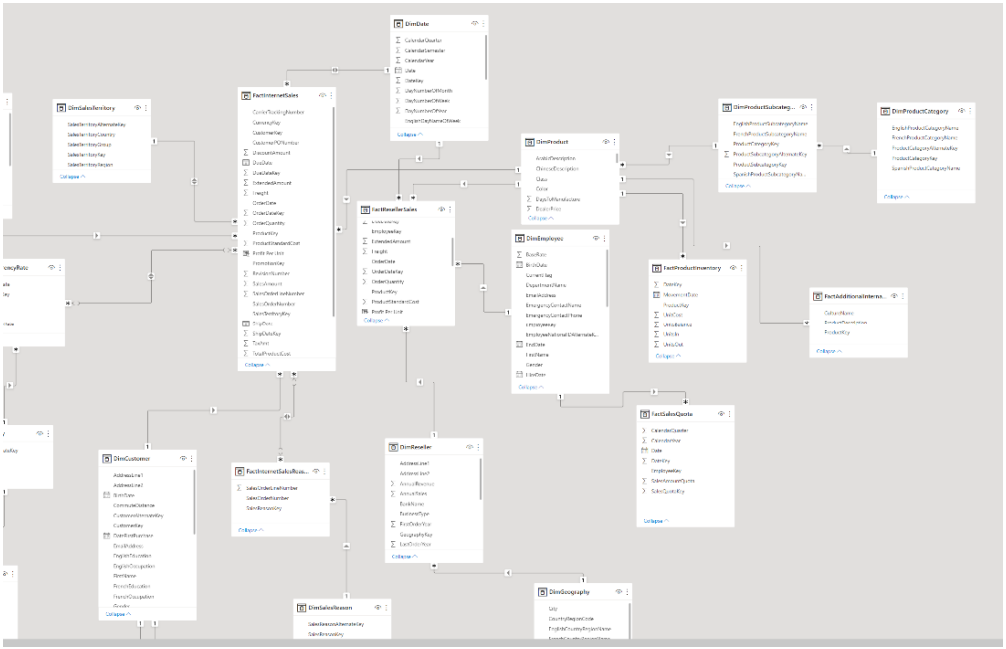
Company Overview

This paper will analyze a fictious bicycle manufacturing company called Adventure Works Cycles (AWC). The dataset is a creation of Microsoft and has been used in many training courses to teach data and analytics. AWC manufactures and sells bicycles, apparel, bicycle components Australia, North America, and Europe. At the time of this analysis, the company has been in operation for approximately three years. The company earns revenue through sales to authorized resellers and directs consumers through their internet website. AWC has produced significant revenue and is seeking to improve profits using analytics.

The Dataset

The AWC dataset is quite impressive, boasting twenty-eight tables and over 100K rows of data. The dataset initially existed in a BAK file format and is a relational database which was restored in an instance of SQL Server Management Studio (SSMS). Once the dataset was restored in SSMS, Microsoft Power BI was used to connect to the AWS dataset in SSMS. Once imported to Power BI, the dataset required the use of the snowflake schema, as there are multiple layers of dimension tables such as DimProduct, DimProductCategory, and

DimProductSubcategory. The various dimension tables contained primary and foreign keys that were used to create relationships with the fact tables. This posed difficulties for the data model due to limitations of Power BI that would not allow multiple active relationships on the same primary key. These relationships would be inactive, therefore if the developer needs to create an active relationship through coding in the Data Analysis Expression language.



Overall, this is a large dataset, which has endless analytics possibilities. However, for the purpose of this exercise, the dataset will focus on two Fact tables. The two Fact tables of interest are the Fact Internet Sales (FIS) and Fact Reseller Sales (FRS). The FIS table captures transactions that occur through AWC's internet sales operations. Fields in the FIS table include but are not limited to Sales Amount, Total Product Cost, Order Date, and Sales Territory. The FRS table is essentially a mirror of the FIS table; however, FRS captures sales transactions with resellers. As with FIS, this study will focus on FRS' Sales Amount, Total Product Cost, and Order Date fields. The analysis of the two fact tables will be centered on revenue and profits.

Descriptive Analysis

This exercise began with an analysis of the data using descriptive analytics. With descriptive analytics, the raw data was transformed into easily understood insights. These insights seek to answer the question: “What happened” (Conrad, 2022). The descriptive analytics portion of this study will methodically evaluate the information derived from the dataset to gain insights from the past performance of the company. The time series for this study spans from Jan 2011 – Jan 2014. The descriptive analysis in this study will be the basis for the machine learning model and the strategic recommendation that will be provided. In this exercise, the descriptive analysis will be a comparison of quantity of units sold and profit generated via internet sales vs. sales to resellers.

The first step in this descriptive analysis was to understand the business problem (Morris, 2022). The business problem at AWC is that while the company is generating substantial revenue, they are not generating sufficient profit given revenue. The next phase was to create a visual that depicted the combined overall performance of the company, depicted in figure 2.

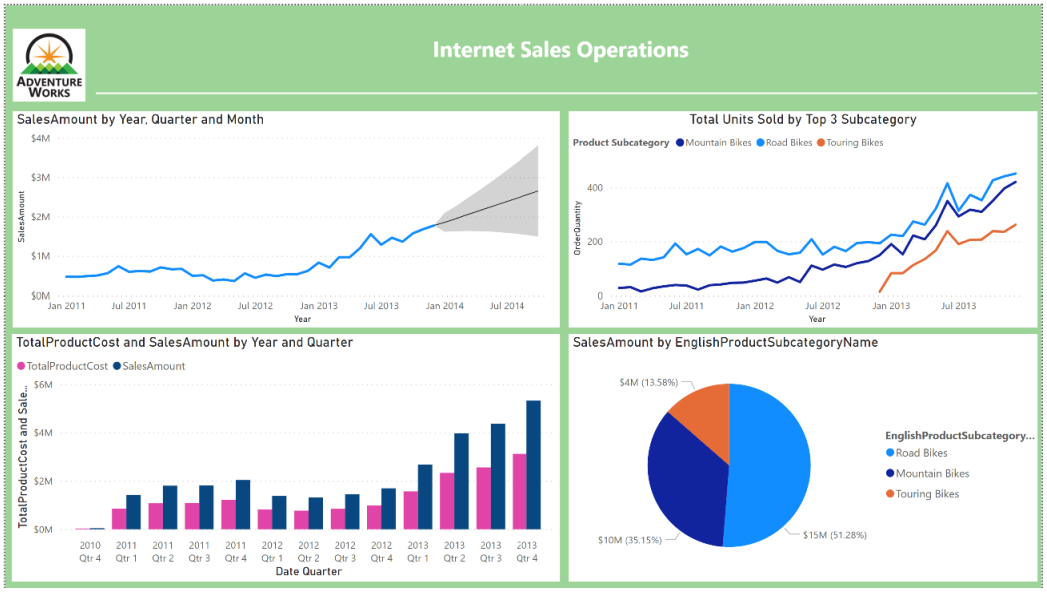
In the first quadrant of the dashboard, we can see a line chart showing the total units sold for the period covered. Additionally, the chart has a trend line, and the grey area is the estimate forecast with min and max estimated units. One look at the units sold line chart and we can observe that the company's units sold exhibit a cyclic trend. A cyclic trend occurs when there is a general up and down trend of demand that spans three years (Doane & Seward, 2019). The cyclic trend typically follows the life of the product and replacement cycles. The major descriptive analytic takeaway from the demand chart is that demand has been cyclic and flat overall. Additionally, forecast growth in demand is likely to remain flat.

Next, moving to the chart in the top right corner of the dashboard, we see a three-line chart depicting demand for the three product subcategories in the dataset. The top three selling product subcategories are road bikes, touring bikes, and mountain bikes. This chart enables the analyst to observe which products have the highest demand, and the potential trend. The important thing to note about the demand by product subcategory chart is that bikes are the top-selling products in AWC, and the demand therein has not exhibited significant growth overall. Furthermore, AWC's road bikes segment has experienced a decline in demand since the inception of the touring line in 2013.

The next chart, in the bottom left quadrant, is a comparison of sales revenue and profit. This chart is very telling, as the observer can see that AWC revenue has experienced an overall growth trend that has exceeded $5M. However, the bar chart portion shows a significant gap between revenue and profit. AWC's overall profit margin has not exceeded 10% throughout this period. Based on the profit-revenue chart, the company has determined that revenue has been strong, and profit has been weak.

The final chart of the first dashboard is the revenue from the top three subcategories chart. This chart shows that road bikes and mountain bikes have generated the most revenue during the period. Touring bikes are new to the product lineup and have only been in service since Jan 2013. Now that the overall company performance has been analyzed, the study will separately analyze the internet sales and reseller sales segments.

The Internet Sales Operations chart is a four-chart dashboard that provides greater granularity in the AWC descriptive analysis. The study will now deep-dive internet sales. In the first chart, the observer will view a line chart that depicts revenue for the period covered. One sees that internet sales had been relatively flat for two years, then experienced nearly 100% growth from Jan 2013 to Jan 2014. Additionally, we see that internet sales are forecast to continue robust growth for at least the next 10 months.



The next chart in the internet sales dashboard is the top three demand by product subcategory. In this chart we are able again to see that AWC's internet sales demand has

experienced steady growth during this period. Each of the top three product subcategories has experienced growth and appears to continue a growth trend.

The next chart on the internet sales dashboard, in the bottom left quadrant, is the total product cost to sales amount. This visual allows the observer to make a general comparison between revenue and product costs. We can see that total product costs have reached 50% of sales revenue.

Moving on to the last visual in the internet sales dashboard, we see the internet sales top three selling product subcategories. Again, this chart shows the strong revenue in all three product subcategories. The touring category is less than the mountain bike and road bike categories, however touring has been in service just one year, and we can observe from the line chart above that touring will continue a growth trend.

The next dashboard to analyze is the Reseller Sales Operations dashboard. Reseller sales are AWC's sales to authorized dealers and retailers. AWC participates in an authorized dealer program which is an optional program that enables AWC to sell their products in a retail business. The company has established agreements with various wholesalers and distributors as well to move AWC products to 3rd party retailers. Historically, utilizing wholesalers was beneficial because it incentivized wholesalers to sell products only to those companies that want to acquire your product (Trackstreet, 2018).

The first chart of the reseller sales dashboard shows revenue over the period analyzed. The analyst can quickly see that reseller sales have been cyclic. Overall, reseller sales have been flat, displaying little growth. Furthermore, reseller revenue is expected to drop slightly, then remain flat.

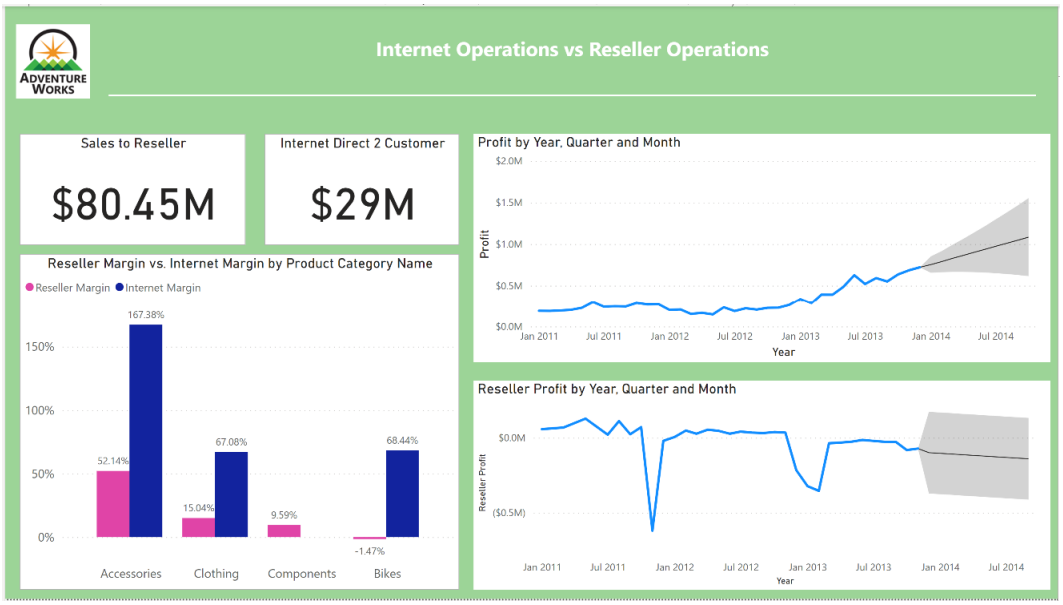
Moving to the next chart, in the reseller sales operations, is top three demand by product subcategory. This chart is quite telling because we see that sales to resellers have been cyclic and have declined since Fall 2012. If you recall the internet sales top three demand by product subcategory experienced demand growth in all three product subcategories.

The next chart in the resellers' sales dashboard is the revenue by month column chart. Here we can observe that reseller revenue comprises a sizable portion of AWC's revenue, however, has experienced a decline in sales in 2013.

The final chart of the reseller sales dashboard shows the sales revenue by the top three subcategories. This chart further shows that sales revenue to resellers has been greater than internet sales over the same period.

The next dashboard in the descriptive analysis is a side-by-side comparison of reseller sales vs internet sales. The metric cards show that revenue from reseller operations is $80M vs $29M for internet sales. Initially, one might assume that reseller operations outperform internet operations. However, the column chart below the metric cards tells a different story. The column chart is a side-by-side depiction of reseller sales and internet sales profit margin. The analyst can

quickly observe that internet sales profit margins are significantly higher in all product categories than those of reseller sales. Furthermore, the internet sales profit margin in the Bikes product category, AWC's top producer, is 69% higher than reseller sales in the same category. The last two-line charts are depictions of internet sales profit and reseller sales profit. The top line chart, internet sales profit shows a growth trend since Jan 2013, and is expected to continue growth pattern. Next, the reseller sales profit depicts an operation that has mostly operated at a loss. Additionally, reseller sales are expected to continue to operate at a loss for at least the next 10 months. Now that the descriptive portion of the study is complete, the analysis will move to predictive analytics.



Predictive Model

Predictive analytics is a tool that can be employed to predict outputs based on inputs (Mason, 2022). Outputs of the predictive model could be used in an entity's decision-making process. The predictive model takes historical data as its input, then employs machine learning to create visualizations that depict trends and/or patterns. Predictive analytics is not a new concept;

however, it was not until the 1940's that computational power was used to build predictive models (Mason, 2022). Additionally, predictive analytics is a global market that is estimated to reach nearly $11B in 2022. Furthermore, predictive analytics enables the firm to transform data into insights that propel data-drive decision making.

The predictive model in this study was built using Python and Juptyer Notebooks. The first step in the predictive analysis was to determine desired functionality of the predictive model, which was to create a model that could predict profit given certain inputs. Next, a simple correlation matrix was created and found that there was collinearity between the variables units sold, revenue, and total product cost. Through dimension reduction, the dataset was reduced to just one independent variable, units sold, and the dependent variable is internet sales profit. The dataset was then grouped by month and summed. Grouping the dataset by month served to smooth the dataset to enable more effective analysis.

Once dimension reduction was complete, units were assigned to an X variable and internet sales profit was assigned to the y variable. The X and y variables were then split into training and test arrays with a test size of 20%. The X and y training variables were then used to fit the data to a linear regression model with Python's Sklearn library. With the data fit to the model, the first predicted model in the study was created the X test array and assigned to a variable, y-predict. The predicted data was then loaded to a Pandas data frame as depicted in table below.

	X	Actual	Predicted
0	144	188040	191856.0
1	4399	531422	560905.0
2	150	193077	192377.0
3	313	225360	206514.0
4	1662	351427	323517.0
5	4087	437778	533844.0
6	269	204010	202698.0
7	5224	739506	632460.0

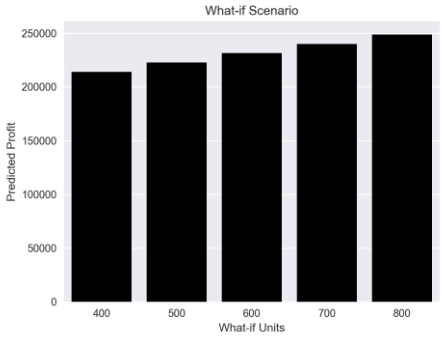
In the chart above we see three variables depicted. The first variable, X, is derived from the X test array and is the independent variable in this predictive model. The next column, actual, is the y test array and is the actual profit from the internet sales profit variable. Finally, the predicted column is the predicted output of the model. A high-level review of the table shows that the model can predict profit based on units of demand. Next, the score function was executed on the X train and y train data and revealed that the model was 79.4% accurate.

Next, the predictive model needs to be repeatable, so a function was created to ensure reusability and portability. The function takes a scalar or list of desired units of demand as an argument, then returns predicted profit.

```
def predictor(nums):
    try:
        pred = regress.predict(np.array([nums]).reshape((-1,1)))
    except:
        print("Exception thrown, incorrect data type entered.")
    return pred
```

As one can see in the snippet above, the predictor function takes a single number or a list unit of demand as its argument, enters the argument into a NumPy array, reshapes the array and returns and array predicted profit given number of units. This process is depicted in the snippet below.

```
sns.barplot(x = what_if,y=result,color='black')
plt.ylabel('Predicted Profit')
plt.xlabel('What-if Units')
plt.title('What-if Scenario');
```



In the above snippet we see an array of units to produce is instantiated in the variable what if and inserted into the predictor function and assigned to the variable result. Next, the what if variable and the result variable is used to create the What-If Scenario column chart above. We can see a general increase in profit as units in demand increases.

Prescriptive Model

Thus far in this study, we have conducted a descriptive analysis that told us that internet sales profit margins had been greater than reseller sales profit margins. Next, the predictive model told us that if demand is X, profit would be y with 79% accuracy. With the first two analyses completed, and predictive model functioning properly, the study moved to the prescriptive model. The prescriptive model answers a different question than the descriptive analysis and prescriptive model. The prescriptive model seeks to answer the question "What actions should be taken" (Lepenioti, et. al., 2019). Lepenioti et. al. contends that much of the analytics work conducted in the business space focuses on descriptive and predictive analytics through data mining and artificial intelligence. They further state that prescriptive analytics is indeed a much less developed discipline than descriptive and predictive analytics space. However, we must not mistake prescriptive analytics immaturity with lack of importance.

Prescriptive analytics is the next evolution of data analytics maturation and enables companies to lean forward and optimize business performance (Lepenioti, et. al., 2019). While prescriptive analytics is indeed the most complex form of analytics, the discipline creates the most value for the firm through enhanced business intelligence.

The prescriptive model in this study is like the predictive model and uses the same libraries and methodologies. The prescriptive model was used to create a function that took profit target as its argument, then returned the number of units of demand to meet that target. An example is depicted in the table below.

	X	Actual	Predicted
0	188040	144	382.0
1	531422	4399	3526.0
2	193077	150	429.0
3	225360	313	724.0
4	351427	1662	1878.0
5	437778	4087	2669.0
6	204010	269	529.0
7	739506	5224	5431.0

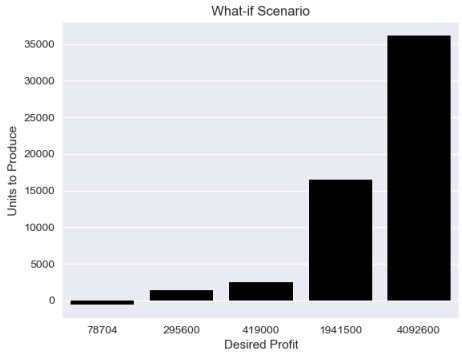
The X variable in the table above is the X test variable internet sales profit test – train data. In the X column we see a list of profit results from various months in the AWC time series. The actual column is the actual number of units of demand in the dataset, and the predicted column is the number of units the model prescribes to reach the target profit. The prescriptive model was scored with the score function and found to be 79% accurate. Once the test data was executed to validate the model, a function was created, like the predictive model, to ensure reusability.

With the prescriptive model functioning properly, a list of target profit was created and executed in the prescriptive model as depicted in the snippet below. The output was again

assigned to an array and used to create the column chart. The prescriptive model in this study is 79% accurate, and could benefit from further optimization, it certainly demonstrates that an analyst can indeed create a working prescriptive model with Python and Jupyter Notebooks. This model is an example of prescriptive analytics because it takes a target profit and prescribes a certain number of units to sell to reach the target profit.

```
what_if = [419000,1941500,4092600,78704,295600]
result = predictor(what_if)
what_if = np.reshape(what_if,len(what_if))
result = np.reshape(result,len(result))
```

```
sns.barplot(x = what_if,y=result,color='black')
plt.ylabel('Units to Produce')
plt.xlabel('Desired Profit')
plt.title('What-if Scenario');
```



Recommendation

The descriptive analysis revealed several key facts about the AWC sales performance. The first is that AWC has produced significant revenue of more than $100M in just three years of operations. However, the descriptive analysis also revealed that while revenue is significant, profit has not performed well given sales revenue. Furthermore, the internet sales and reseller sales comparison revealed that internet sales profit margins are significantly higher than reseller sales profit margins. The first recommendation is to utilize the predictive and prescriptive models created in this study to build other models such as sales, marketing, and manufacturing. Both models are lightweight, flexible, and repeatable and can be adapted to most situations.

Next, the company must explore a shift in strategy from reseller sales focus to direct to consumer via internet sales operations. One example of the company shifting focus from resellers to direct to consumer is Nike apparel manufacturer. The company announced a shift from focusing on resellers to direct to consumer in 2017 (Nike News, 2017). The shift would reduce the number of authorized resellers from 30,000 to just 40 (Danzinger, 2018). This strategy has proven fruitful for Nike, with 22Q3 revenues up 5% at $10.9B (Nike News, 2022).

Conclusion

This study has proven that it is indeed possible to utilize descriptive, predictive, and prescriptive analytics in a flexible and repeatable model. The predictive and prescriptive models in this study are indeed raw and would benefit from further optimization. This optimization could include integration into an artificial neural network to increase accuracy. Further research and testing of the two models is recommended.

References

JIT Just-in-Time manufacturing. (n.d.). www.ifm.eng.cam.ac.uk. https://www.ifm.eng.cam.ac.uk/research/dstools/jit-just-in-time-manufacturing/#:~:text=JIT%20is%20a%20Japanese%20management

How to Detect and Correct Multicollinearity in Regression Models. (n.d.). Engineering Education (EngEd) Program | Section. https://www.section.io/engineering-education/multicollinearity/

Nau, R. (2019). *What's a good value for R-squared?* Duke.edu. https://people.duke.edu/~rnau/rsquared.htm

PennState: Statistics Online Courses. Retrieved March 26, 2022, from https://online.stat.psu.edu/stat508/lesson/2/2.2

What is Descriptive Analytics? A Definition. (n.d.). GetApp. Retrieved March 26, 2022, from https://www.getapp.com/resources/descriptive-analytics-definition/

NetSuite.com. (n.d.). *Descriptive Analysis Defined*. Oracle NetSuite. Retrieved March 26, 2022, from https://www.netsuite.com/portal/resource/articles/erp/descriptive-analytics.shtml#:~:text=%20Five%20Steps%20in%20Descriptive%20Analytics%20%201

Schydlowsky, A. (n.d.). *TrackStreet*. Trackstreet. Retrieved March 26, 2022, from https://www.trackstreet.com/authorized-dealer-program/https://www.trackstreet.com/authorized-dealer-program/

What Is Predictive Analytics? (2021, July 7). William & Mary. https://online.mason.wm.edu/blog/what-is-predictive-analytics

Lepenioti, K., Bousdekis, A., Apostolou, D., & Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, *50*, 57–70. https://doi.org/10.1016/j.ijinfomgt.2019.04.003

Nike. (2017). *NIKE, Inc. Announces New Consumer Direct Offense: A Faster Pipeline to Serve Consumers Personally, At Scale*. Nike News. https://news.nike.com/news/nike-consumer-direct-offense

Danziger, P. N. (n.d.). *Nike's New Consumer Experience Distribution Strategy Hits The Ground Running*. Forbes. Retrieved March 26, 2022, from https://www.forbes.com/sites/pamdanziger/2018/12/01/nikes-new-consumer-experience-distribution-strategy-hits-the-ground-running/?sh=2fab1f29f1d0

NIKE, Inc. Reports Fiscal 2022 Third Quarter Results. (n.d.). Nike News. Retrieved March 26, 2022, from https://news.nike.com/news/nike-inc-reports-fiscal-2022-third-quarter-results

Doane, D. P., & Lori Welte Seward. (2019). *Applied statistics in business and economics*. Mcgraw-Hill/Irwin.

Nguyen, Giang & Kedia, Jai & Snyder, Ryan & Pasteur, R. & Wooster, Robert. (2013). Sales Forecasting Using Regression and Artificial Neural Networks.