*DA 101*

Final Project

*Reams, Quinn*

*Penn State University*

Business Understanding

The objective of this analysis is to investigate and predict the gross drug costs for patients based on various factors such as risk score, age, gender, and specialty drug usage. In other words, the goal is to identify the primary variables influencing gross drug costs and deliver insights to better predict the gross drug cost and reduce expenses.

**Predictions about gross drug cost:**

* RiskScore will affect the gross drug cost positively, with higher risk scores driving gross drug cost higher.
* Age will affect gross drug cost positively, with older ages driving gross drug costs higher.
* SpecialtyDummy will affect the gross drug cost positively, if that patient needs specialty medicine.
* Gender will have a minimal impact on gross drug cost.

**Data Understanding**

**Description of data variables:**

* **GrossDrugCost** – The total dollar amount of drug related expenses incurred by a member during the specified month. This is a continuous numerical variable.
* **NLISDummy** – A dummy variable indicating whether a member is classified as non-low income (1 = Non-Low Income, 0 = Low Income). NLISDummy is classified as a nominal, dummy, categorical variable.
* **LISCHOSERDummy** – A dummy variable indicating whether a member opted for a specific drug plan (1 = Chose Plan, 0 = Auto-Assigned Plan). LISCHOSERDummy is a nominal, dummy, categorical variable.
* **RiskScore** – A continuous numerical variable, based on previous government data, reflecting how sick someone is, with higher scores signifying sicker members.
* **SpecialtyDummy** – A dummy indicator indicating whether a member receives specialty drugs (1 = Yes, 0 = No). This is a nominal, dummy, categorical variable.
* **AdjudicationDays** – The number of non-holiday workdays within the corresponding month. This is a discrete numerical variable.
* **Age** – A discrete numerical variable indicating the age of the corresponding member.
* **Gender** – A dummy, nominal, categorical variable indicating the gender of the member (1 = Female, 0 = Male).
* **FralityDummy** – A dummy, nominal, categorical variable that indicates if the government has a record on whether or not the member is frail (1 = Frail, 0 = Not Frail).
* **HospiceDummy** – This is a dummy, nominal, categorical variable that indicates if the member is receiving hospice care (1 = Receiving Hospice Care, 0 = Not Receiving Hospice Care).
* **InstitutionDummy** – A dummy variable that indicates whether or not the member is receiving institutionalized long term care (ie., Nursing Home, Hospital) (1 = Receiving, 0 = Not Receiving).
* **ESRDDummy** – A dummy variable that indicates whether the member is receiving end-stage renal disease treatment (ie., end-stage kidney disease) (1 = Is Receiving, 0 = Is **Not** Receiving).

**Data Modeling and Evaluation**

**Descriptive Analytics**:

**Figure 1 - Descriptive Analytics Table:**

A screenshot of a computer

Description automatically generated

**Figure 2 - Box Plot without Outliers:**

A graph with a blue square and a black dot

Description automatically generated

\**Tool Used: Minitab.*

**Figure 3 – Bar Chart of Gross Drug Cost by Gender:**

A green and blue bar

Description automatically generated

*\*Tool Used: Tableau.*

**Difference of Gross Drug Cost by Gender(diagnostic analysis)**:

The bar chart illustrates that the average gross drug cost by gender shows only a minimal difference, with males averaging $533.85 and females averaging $545.66. The Pearson correlation coefficient (*R* = 0.004) and the 95% confidence interval for the difference in means (-0.007, 0.015), calculated using Minitab, confirm that there is no statistically significant relationship between gender and gross drug cost. This shows that gender has negligible a impact on gross drug cost.

**Figure 4 - Correlation of Risk Score and Gross Drug Cost**:

A graph of a number of blue dots

Description automatically generated

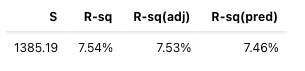
\**Tool Used: Minitab.*

The relationship between risk score and gross drug cost shows a weak positive correlation (*R* = 0.274, 95% CI: 0.264-0.285). Most data points cluster below $15,000 in drug costs and under 3,000 in risk score, with scattered outliers extending to $60,000 in drug cost. While drug cost tends to increase slightly with risk score, the weak correlation suggests that risk score is not a reliable predictor gross drug cost.

**Predictive Analysis Knowledge**:

Model 1: GrossDrugCost = B0 + B1 \* RiskScore + ε *\*Tool Used: Minitab.*

A screenshot of a calculator

Description automatically generated

* **Model Type:** Simple Regression Model
* **Significance:** Highly Significant (P–Value = 0.000).
* **Impact per Unit Increase:** 1.11$ for every one unit increase.
* **Adjusted R-squared interpretation:** The adjusted R-squared is 7.53%, indicating that about 7.53% of the variability in gross drug cost is explained by risk score. While the relationship is statistically significant, other factors contribute to gross drug cost variability.

Model 2: GrossDrugCost = B0 + B1 \* RiskScore + B2 \* Age + B3 \* Gender + ε

*\*Tool Used: Minitab*

A screenshot of a graph

Description automatically generatedA screenshot of a computer

Description automatically generated

* **Model Type:** Multiple Regression Model
* **Model Comparison to Model 1:** In this Model we have multiple variables contributing to the model, adding Age and Gender.
* **Comparison of Adjusted R Squared to Model 1:** Model 2 (7.76%) is slightly higher than Model 1 (7.53%), which indicates that model 2 contributes slightly more variability than Model 1.
* **Conclusion of Comparison:** The slight increase can be explained by the added variables, though the difference is minimal (the contribution of gender wasn’t even statistically significant).

Model 3: GrossDrugCost = B0 + B1\*RiskScore + B2\*SpecialtyDummy + B3\*Age +  B4\*Gender + ε *\*Tool Used: Minitab.*

A screenshot of a graph

Description automatically generatedA screenshot of a phone

Description automatically generated

* **Significant Predictors:** SpecialtyDummy, Age, and RiskScore.
* **Comparison of Models:** Model 3 is the strongest of the three for predicting gross drug cost, with an adjusted R-squared of 10.13%, the highest between the models. The inclusion of SpecialtyDummy (Coef = 457.7, P < 0.001) substantially increases the model’s explanatory power, contributing an improvement that the other two models lack. This variable explains the most variability in gross drug cost, making Model 3 the most reliable for prediction.

**Key Findings and Recommendation**:

* Two of our assumptions were accurate, (RiskScore and SpecialtyDummy)!
* Gender did not play as significant of a role as we anticipated.
* Age impacted the change in a negative relationship instead of a positive.
* I would recommend using Model 3. Despite it being more complex, it explained the highest variability.

**Telling the Story**:

* **$1.11**: This is the amount we predict gross drug score will go up per unit based on RiskScore. It’s *risky* business not being able to predict expenses. That’s profit left on the table and unaccounted for! That is what brings us to our first model, a simple regression model to predict the gross drug cost using the risk score.
* **Age Matters**: In Model 2 we try to also predict gross drug cost using age. We found a little success in this, as age showed us that with every year, price goes down $6.06.
* **Gender Differences**: Testing and hypothesis don’t come without their surprising finds. Gender ended up not being a very strong signifier, or even significant at all, when we expected it to be.
* **$467.70**: That big find is the amount of difference a specialty drug can make on a cost. Adding this in the prediction for Model 3 was our absolute biggest find since risk score!
* **Model 3**: Model 3 is the model I would recommend of the three. It has proven to be the best predictor by a margin of almost 2.5%!