**BUAN 6383 - Project 2**

Rhett James Gambrell

Parth Nitin Kahane

Munsif Junaid

Harish Kumar Sarathi

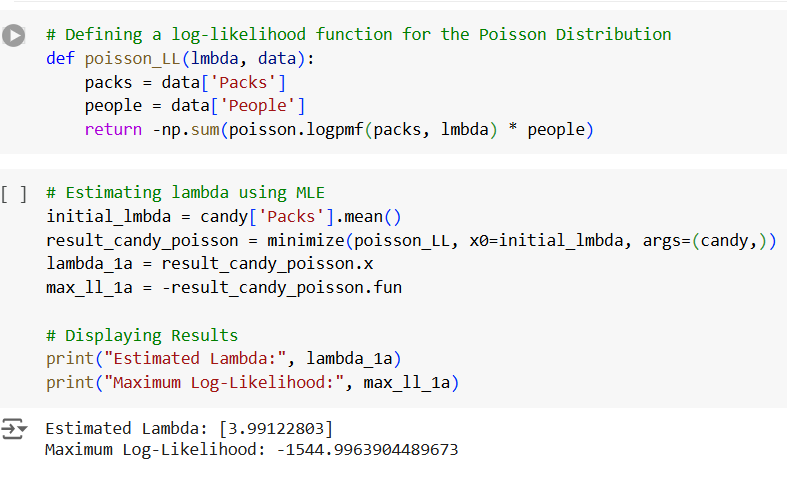
Vijay Venkatesan

**PART - 1**

Question-1)  
A) The lambda value for the Poisson model after optimization is 3.99122803 & the maximum Log likelihood function value for the Poisson model is obtained as -1544.9963904489673

AIC: 3091.9927808979346

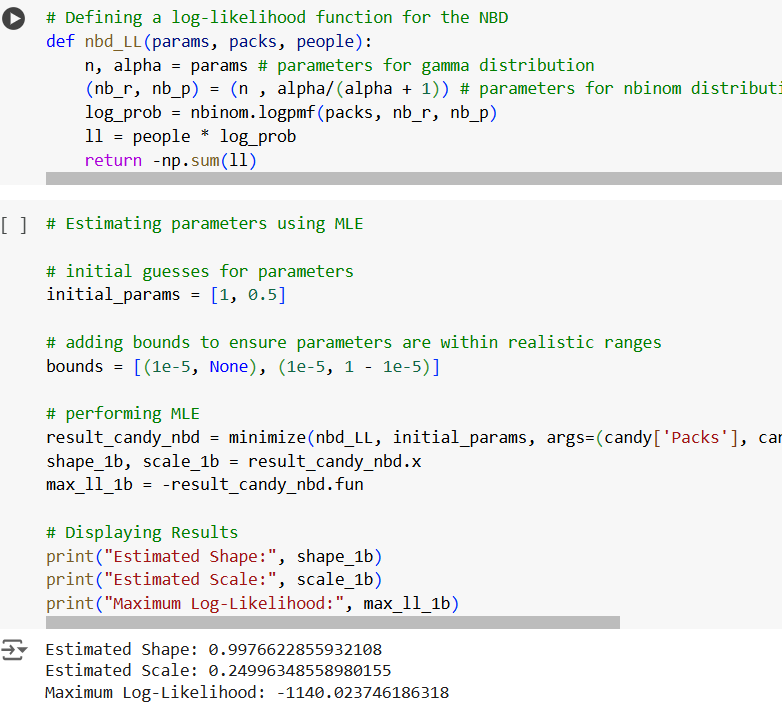
BIC: 3096.115273707449



B) The Alpha value for the NBD model is 0.24996348558980155, Shape parameter value for NBD model is 0.9976622855932108 and the Log likelihood function value for NBD model after optimization is -1140.023746186318

AIC: 2284.047492372636

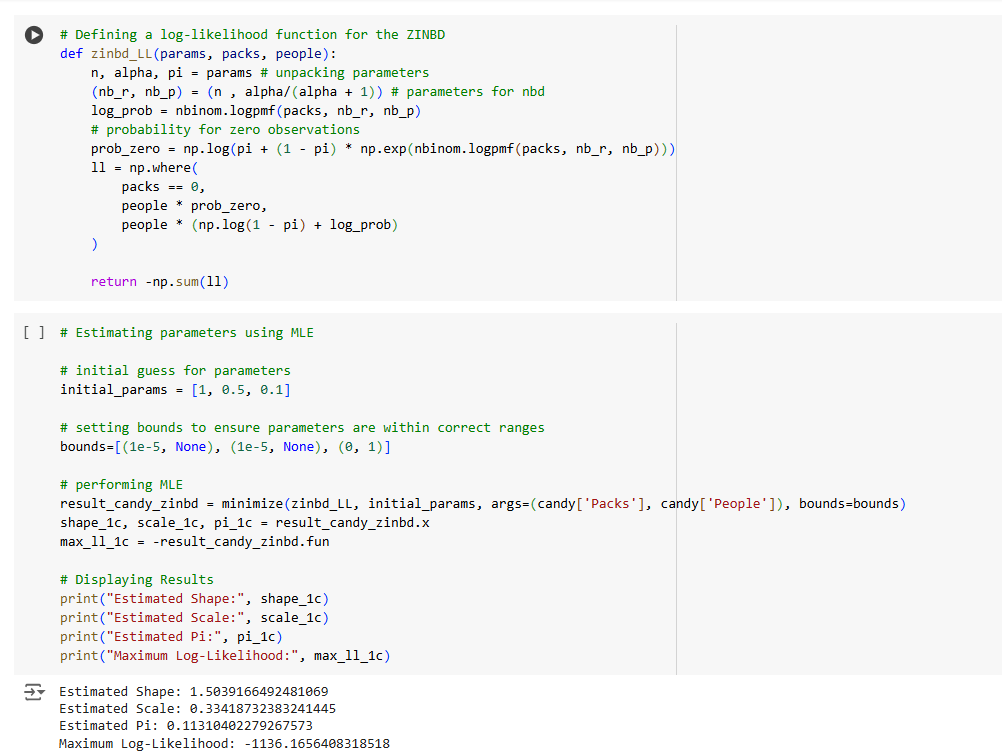
BIC: 2292.292477991665



C) The Scale value for the Zero-inflated NBD model 0.33418732383241445, shape parameter value for Zero-inflated NBD model 1.5039166492481069, Pi value for Zero-inflated NBD model 0.11310402279267573, Maximum Log likelihood value for Zero inflated NBD model after optimization -1136.1656408318518,

AIC: 2278.3312816637035

BIC: 2290.6987600922466



D)

i) 2 - Segment

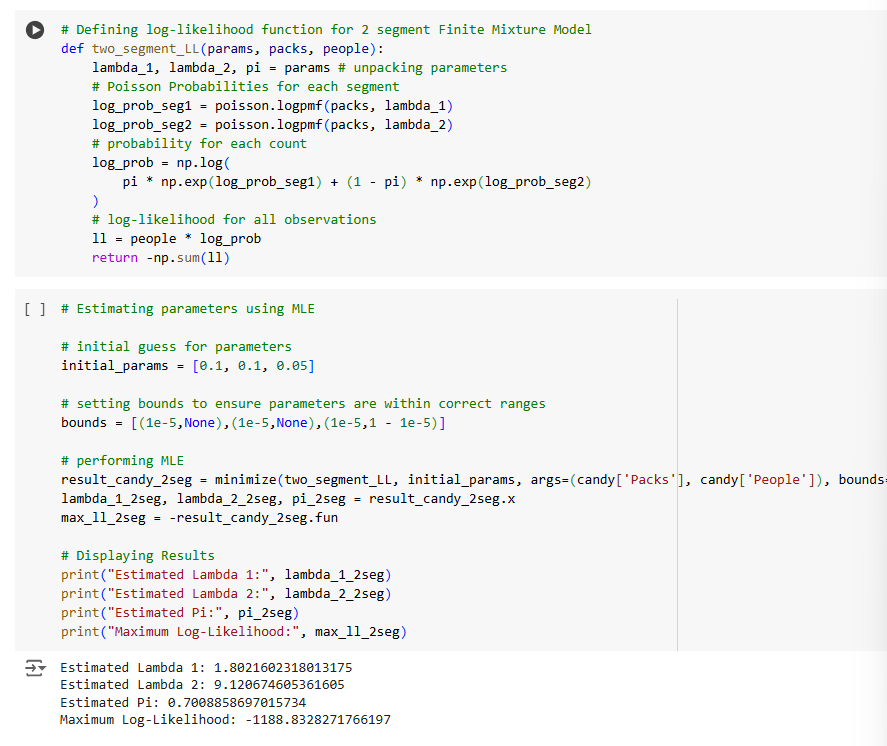
Estimated λ1 value for two-segment 1.8021602318013175, λ2 value for two-segment 9.120674605361605,

π1 value for two-segment 0.7008858697015734, π2 = (1- π 1) value for two-segment 0. 2991141302984266

Maximum Log likelihood value for 2-segment finite mixture model after optimization -1188.8328271766197

AIC: 2383.6656543532395

BIC: 2396.0331327817826



ii) 3 - Segment

λ1 value for three-segment 11.21597748456169,

λ2 value for three-segment 3.4832222009357725,

λ3 value for three-segment 0.29053444936941436,

θ1 value for three-segment - 0.43029411563910414,

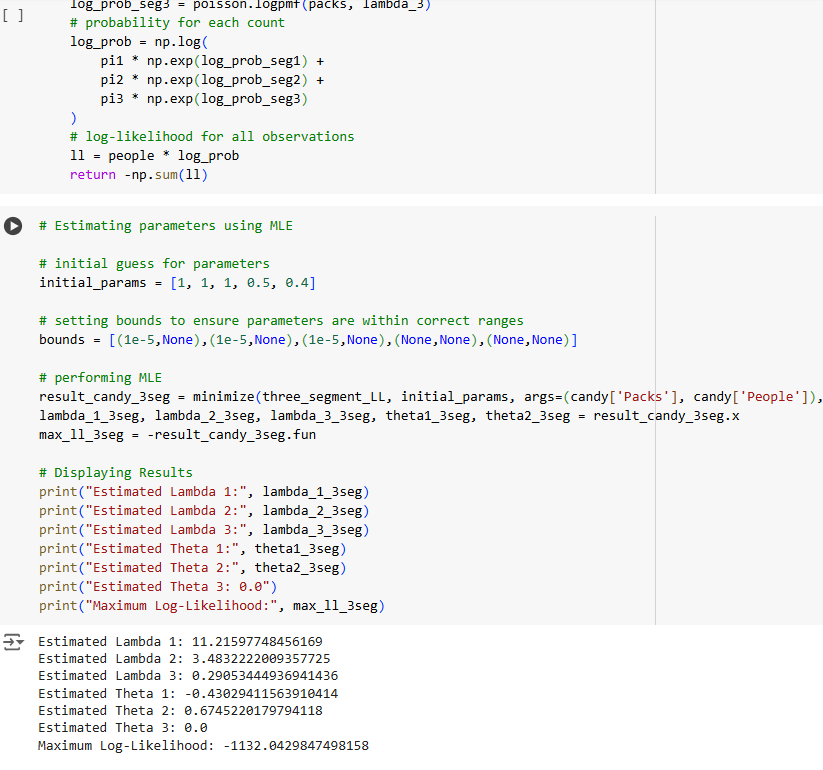
θ2 value for three-segment 0.6745220179794118,

θ3 value for three-segment 0.0

Maximum Log likelihood value for 3-segment finite mixture model after optimization - 1132.0429847498158

AIC: 2274.0859694996316

BIC: 2294.6984335472034



iii) 4 - Segment

λ1 value for three-segment 12.87247290581824,

λ2 value for three-segment 7.418068738487745,

λ3 value for three-segment 3.0019353642157904,

λ4 value for four-segment 0.2047315675306965,

θ1 value for three-segment - 0.8759768180273021,

θ2 value for three-segment - 0.47811261496321306,

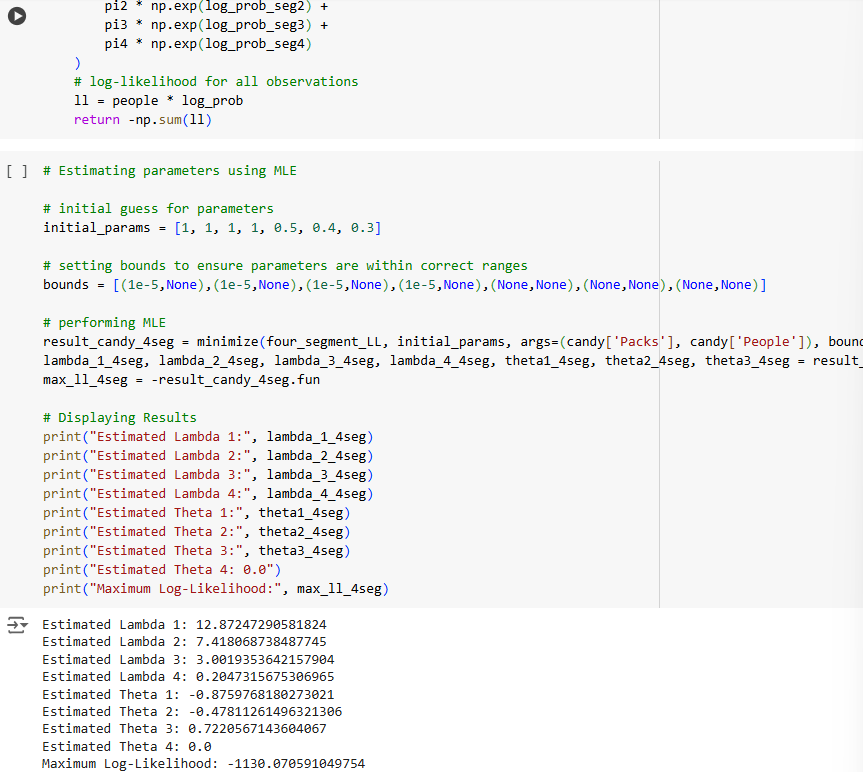
θ3 value for three-segment 0.7220567143604067,

θ4 value for four-segment 0.0,

Maximum Log likelihood value for 4-segment finite mixture model after optimization - 1130.070591049754

AIC: 2274.141182099508

BIC: 2302.9986317661087



Question 2)

A screenshot of a graph

Description automatically generated

Looking at the AIC and BIC scores shows clear differences in how the models perform. The 3 Segment Mixture Model has the lowest AIC, meaning it does a good job of capturing the data without being overly complicated. On the other hand, the Zero-Inflated NBD has the lowest BIC, showing it fits the data well while keeping the model simple.

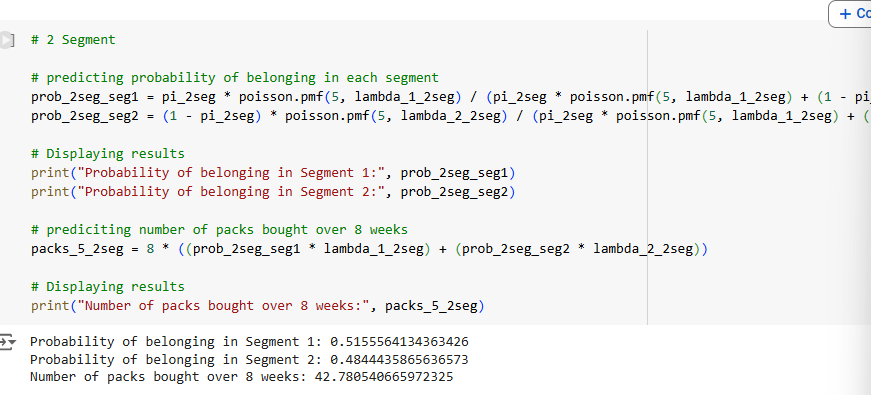
The Poisson model is easy to use but doesn’t handle overdispersion or variation in the data as well as NBD-based models. The 3 Segment Model is great for identifying groups within the data, while the Zero-Inflated NBD balances accuracy and simplicity, making both useful for different purposes.

Question 3)

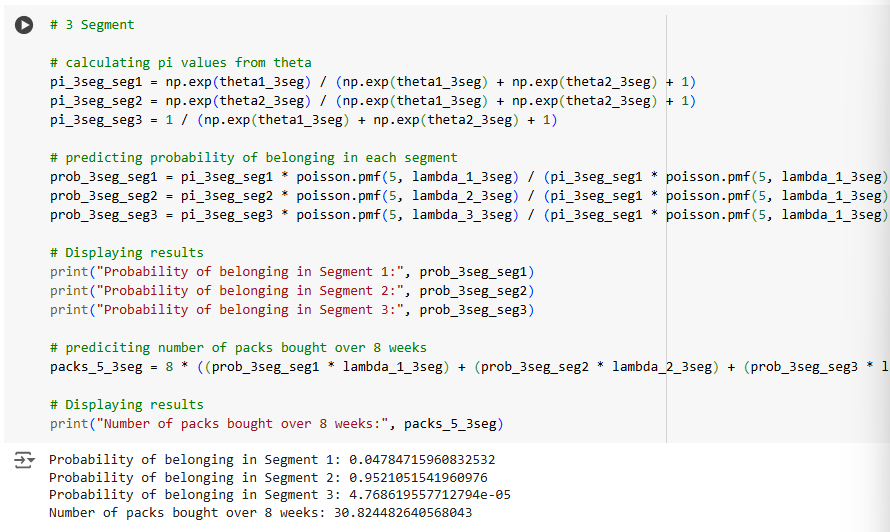
**A) Customer who purchased 5 packs in the past week**

Given that the customer has purchased 5 packs in the last week, the following are the expected values for purchases in the next 8 weeks:

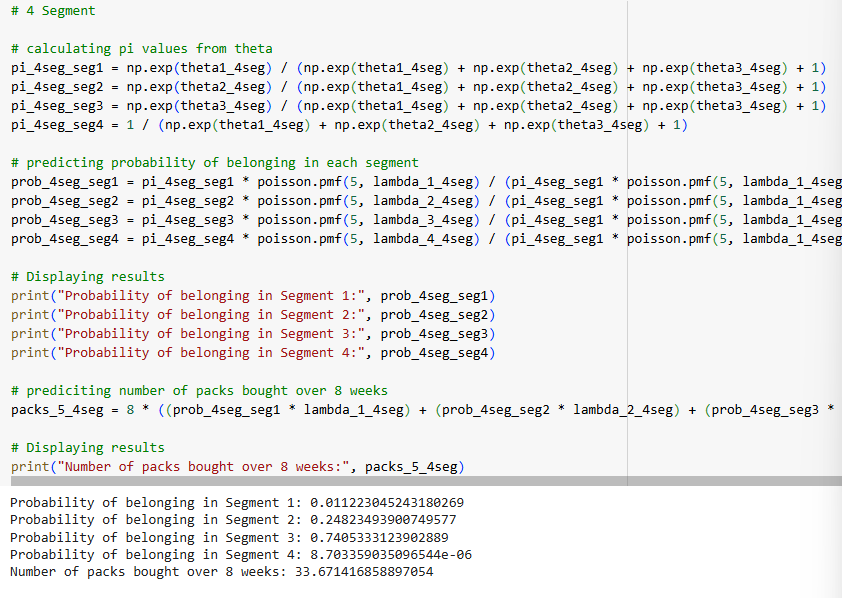
For 2- Segment: # of packs bought over 8 weeks is 42.780540665972325



For 3- Segment: # of packs bought over 8 weeks is 30.824482640568043



For 4- Segment: # of packs bought over 8 weeks is 33.671416858897054

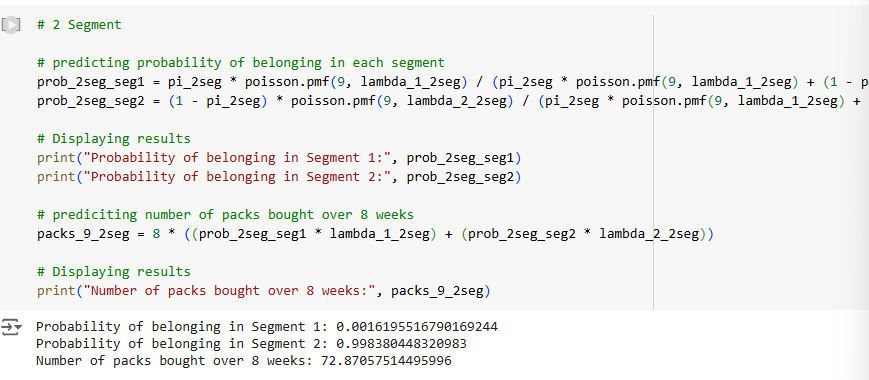


The models suggest the customer is likely to purchase between 30 and 43 packs over the next 8 weeks. The 2-Segment Model gives the highest estimate, while the 3-Segment Model predicts the lowest. This range reflects differences in how each model captures customer behavior. Importantly, all models suggest the customer might not maintain a steady purchase rate of 5 packs per week, as their predictions are slightly lower than 5(packs) X 8(weeks) = 40

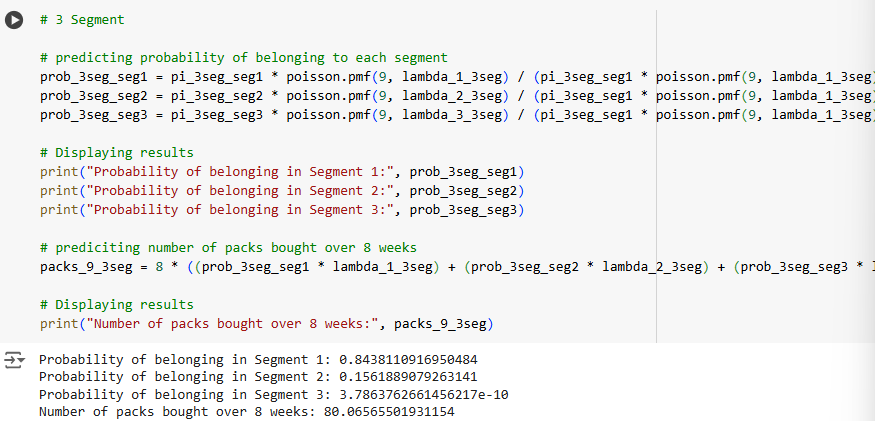
B) **Customer who purchased 9 packs in the past week**

The following is the likelihood of the total amount of purchases the customer will make in the next 8 weeks, given they have made 9 purchases this week, model dependent:

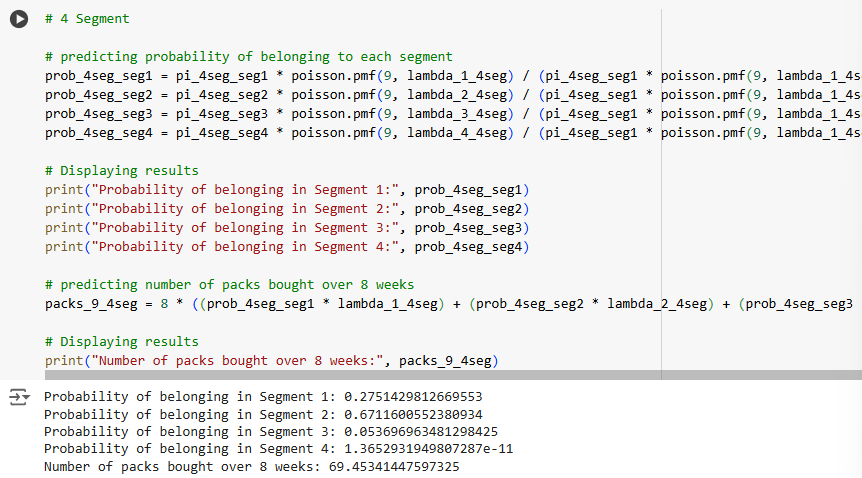
For 2- Segment: # of packs bought over 9 weeks is 72.87057514495996



For 3- Segment: # of packs bought over 9 weeks is 80.06565501931154



For 4- Segment: # of packs bought over 9 weeks is 69.45341447597325



The models predict the customer will buy between ~69 and ~80 packs over the next 8 weeks. The 3-Segment Model gives the highest estimate, while the 4-Segment Model predicts the lowest. The 2-Segment Model prediction is close to a simple calculation of 9 (Packs) X 8 (Weeks) = 72.

These variations suggest the models handle customer behaviour differently, and selecting the most accurate one may require additional domain knowledge about the purchasing patterns.

**PART 2**

**Question 1) Poisson Regression**

For Poisson Regression, the estimated parameters (excluding intercept *const*) are *female, married, kids, prestige* and  *menpubs.*

The maximum log-likelihood comes out to be -1651.0563161667305

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

The analysis identifies key predictors of engagement based on p-values and coefficients. Among the significant predictors, menpubs\text{menpubs} has the largest positive effect, increasing engagement by 0.0255 per unit, making it the most influential factor. In contrast, female\text{female} and kids\text{kids} show meaningful negative effects, indicating potential barriers for women and individuals with children. Married\text{Married} has a moderate positive impact, contributing to engagement but less so than menpubs\text{menpubs}.

The predictor prestige\text{prestige}, while included in the model, is not a significant factor and has minimal impact on engagement. Managerial efforts should focus on leveraging the positive influence of menpubs\text{menpubs} while addressing the challenges associated with female\text{female} and kids\text{kids} to promote greater equity in engagement. Married\text{Married} can be considered a supporting characteristic, while prestige\text{prestige} holds the least importance.

**Predicting number of candidates using Poisson-**

The predicted value for 5 articles is obtained by first calculating the expected article count (λ) for each observation using the model [λ=e^(X⋅params)]. Then, the **Poisson PMF** is used to compute the probability of exactly 5 articles for each observation. Finally, these probabilities are summed across all observations to give the total expected count of candidates with 5 articles. This process ensures the prediction reflects the model's distribution of article counts.

Expected Count of 5 articles = 22.83419923521511

A screenshot of a graph

Description automatically generated A screenshot of a graph

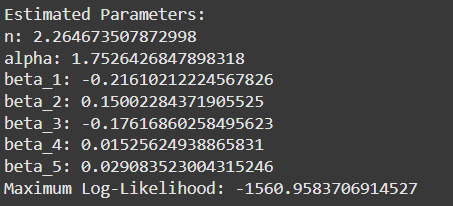
Description automatically generated

A graph with blue and orange bars

Description automatically generated

**Question 2) NBD Regression**

The following displays the estimated parameters-



The analysis identifies clear impacts of the predictors on the log-transformed response variable. β1\beta\_1β1​ has a negative effect, reducing the response variable with each unit increase. This suggests the need to limit its influence or target groups less affected by it. β2\beta\_2β2​, on the other hand, has a positive effect, increasing outcomes and highlighting the importance of focusing on customers with higher β2\beta\_2β2​ values. β3\beta\_3β3​ negatively impacts engagement and may require strategies to address its effects, while β4\beta\_4β4​ and β5\beta\_5β5​ show weaker positive impacts, with relevance primarily in specific contexts.

The most significant predictors are β1\beta\_1β1​ and β2\beta\_2β2​, as their coefficients show stronger impacts. Efforts should center on leveraging β2\beta\_2β2​ to boost outcomes and managing β1\beta\_1β1​ to reduce its negative effects. This approach can help optimize overall results by focusing on the most influential factors.

**Predicting number of candidates using NBD:**

To predict the number of individuals with 5 articles using the Negative Binomial Distribution (NBD), the expected rate exp⁡(β⋅X)\exp(\beta \cdot X) is first calculated for each record. The probability of success pp is then determined as p=α/(α+exp⁡(β⋅X))p = \alpha / (\alpha + \exp(\beta \cdot X)). Using the NBD probability mass function (PMF), the probability of observing exactly 5 articles is computed for each record. Finally, these probabilities are summed across all records to obtain the total predicted count for individuals with 5 articles.

A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

A graph of a number of candidates

Description automatically generated

**Question 3) Zero Inflated NBD Regression**

The model accounts for two sources of zeros in the data: Type-0 candidates (π\pi), who are inherently non-publishers and always have zero articles, and potential publishers (1−π1-\pi), who may have zero articles but follow a Negative Binomial Distribution (NBD) for the number of articles. This dual-source structure explains overdispersion, with the probability for y=0y = 0 combining contributions from both groups, while probabilities for y>0y > 0 arise only from the 1−π1-\pi group.

Using a log-link function, the NBD component models count probabilities based on predictors (XX). The log-likelihood function incorporates probabilities for y=0y = 0 and y>0y > 0, summing them across all records. Parameters (π,n,α,β\pi, n, \alpha, \beta) are estimated through Maximum Likelihood Estimation (MLE) by minimizing the negative log-likelihood, effectively capturing the dual-zero structure and overdispersion in real-world data.

A screenshot of a computer

Description automatically generated

Key takeaways include prioritizing characteristics with significant positive β\beta values to identify candidates likely to become active publishers and addressing significant negative β\beta values to reduce barriers to publishing. Additionally, the parameter π\pi can guide resource allocation, helping determine whether to focus on converting non-publishers into active publishers or improving productivity among potential publishers.

To predict the count of candidates with 5 articles, the model calculates exp⁡(β⋅X)\exp(\beta \cdot X) for each record and derives pp as α/(α+exp⁡(β⋅X))\alpha / (\alpha + \exp(\beta \cdot X)). Using the Negative Binomial PMF, the probabilities for k=5k = 5 are computed as (1−π)⋅nbinom.pmf(5,n,p)(1 - \pi) \cdot \text{nbinom.pmf}(5, n, p). Summing these probabilities across all records gives the total predicted count for 5 articles, integrating contributions from the zero-inflation and NBD components.

A screenshot of a graph

Description automatically generated A screenshot of a graph

Description automatically generated

A graph with numbers and lines

Description automatically generated

**Question 4)**

We evaluate the models based on AIC and BIC.

A screenshot of a number of parameters

Description automatically generated

Based on the calculations considering both AIC and BIC, the **Negative Binomial Regression** is the best performing model.

Based on AIC and BIC calculations, the Negative Binomial (NBD) Regression performs the best overall. The NBD model achieves the lowest AIC and BIC, indicating it provides the best fit for the data. In contrast, the Poisson Regression struggles with overdispersion, as it assumes equal mean and variance, which does not hold in this case. The Zero-Inflated (ZI) NBD model, while accounting for excess zeros, overcomplicates the model and does not lead to better fit, resulting in higher AIC/BIC values.

The key differences arise from how each model handles the data's characteristics. The NBD model effectively captures variability without needing a zero-inflation component, making it both simple and accurate. Meanwhile, the ZI NBD's additional parameters increase the AIC/BIC penalties without improving the model's fit. Overall, the NBD Regression strikes the best balance between simplicity and accuracy, making it the most suitable model for this dataset.

**Managerial Takeaways:**

Identifying key customer characteristics with positive or negative impacts can help tailor strategies for different groups, such as focusing on potential publishers or addressing barriers for those less likely to publish. Analyzing zero inflation has clarified the distinction between customers who are inactive by nature and those temporarily inactive, guiding resource allocation. Balancing model complexity with interpretability is crucial for making actionable decisions, as both need to be considered for effective outcomes. Testing different models reinforced the importance of aligning statistical techniques with real-world contexts to improve decision-making.

**INSIGHTS**

Part 1: The manager can enhance mentorship programs by connecting candidates with highly productive mentors and providing additional support to candidates facing challenges, such as childcare responsibilities or being in less prestigious departments. In the case of marketing candy, the manager should identify high-potential customer segments and design targeted promotions to boost sales. For customers identified as inactive but with potential to return, the manager can implement tailored re-engagement campaigns, such as offering personalized discounts or reminders, to reignite their interest. These actions align resources with data-backed priorities, improving efficiency and results**.**

Part 2: It is very important to understand how different models handle data characteristics such as overdispersion and zero inflation when selecting the right model. Evaluating models using metrics like AIC and BIC helps balance complexity and fit. The NBD Regression proved to be the best fit for this dataset, as it effectively handles overdispersion, making it more robust than both simpler and more complex models. The ZI NBD Regression highlighted the need for model assumptions, like excess zeros, to align with the data characteristics to avoid overcomplicating the model without improving fit.