Databyday - Career Consulting Analysis

Executive Summary

This analysis explores the simulated customer and sales data of *Databyday*, my career consulting business that aims to support individuals in learning the skills of a data analyst to pursue the profession. The primary goals are to:

1. Enhance **customer conversion** across product lines (1-on-1 career sessions, data eBooks, and affiliate marketing courses).
2. Improve **customer retention** and **product performance** through insight-driven strategies.

Key Points:

1. **Demographic and Behavioral Insights**:
    a. **Customer Demographics**: Majority of customers are aged 20-40, with a slightly higher percentage of females, primarily from urban and suburban areas.
    b. **Spending Behavior**: Spend distribution shows two clusters: low spenders (< $30) and high spenders (> $100). Engagement scores are right-tailed, with most users exhibiting low engagement.
2. **Upsell A/B Test Success**:
    a. **Key Result**: The treatment group receiving upsell messages showed a conversion rate of 59.6%, compared to 28.2% for the control group (P-value < 0.0001).
    b. **Recommendation**: Roll out upsell messages and conduct subgroup analyses to optimize strategies for individuals less likely to choose the upsell full product.
3. **Funnel Analysis**:
    a. **Strengths**: High conversion rates from lead to engagement (68.95%) and from engagement to conversion (70.93%) demonstrate effective top- and mid-funnel strategies.
    b. **Opportunities**: Address barriers for the 31% of leads and 29% of engaged users who drop off, potentially testing price sensitivity or marketing messaging strategies.
4. **Product Performance and NPS**:
    a. **Top Products**: The eBook is the most popular product, followed by the data course. The average spend per user is $48.14, with most spending $30 or less.
    b. **NPS Insights**: Scores are consistent (~7), with lower scores in the 50-64 age group. Gathering qualitative feedback could be useful in understanding how to improve satisfaction and NPS.
5. **Predictive Modeling for Retention**:

a. **Model Performance**: Random Forest with Recursive Feature Elimination achieved an ROC AUC of 0.81 and accuracy of 0.75, balancing computational efficiency and strong performance.
b. **Recommendation**: Proceed with Random Forest (RFE) for predicting retention and revisit using XGBoost for datasets with higher complexity.

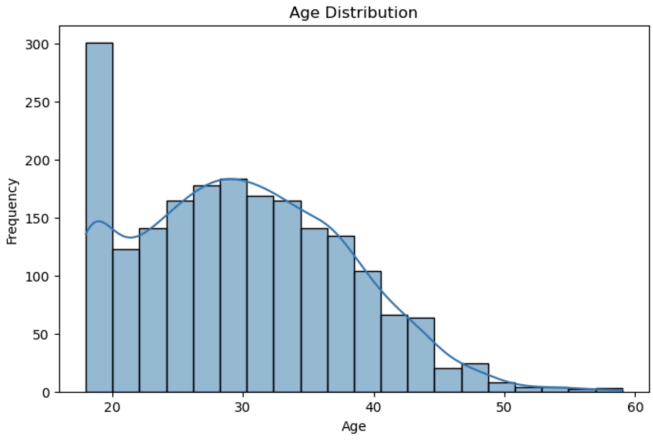
Databyday - Career Consulting Analysis Full Report

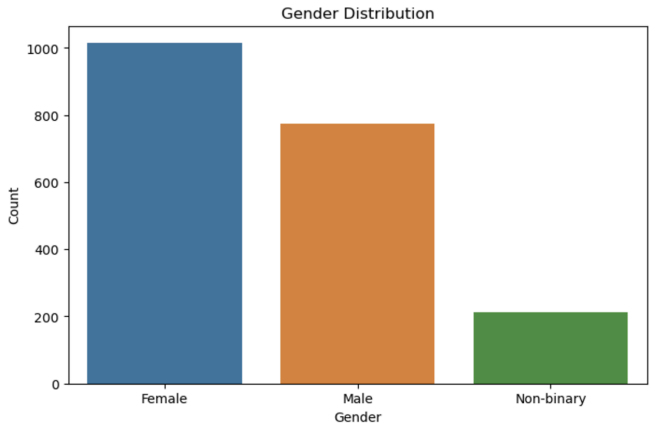
This analysis explores the simulated customer and sales data of *Databyday*, my career consulting business that aims to support individuals in learning the skills of a data analyst to pursue the profession. The primary goals are to:

1. Enhance **customer conversion** across product lines (1-on-1 career sessions, data eBooks, and affiliate marketing courses).
2. Improve **customer retention** and **product performance** through insight-driven strategies.

Demographic and Behavioral Insights

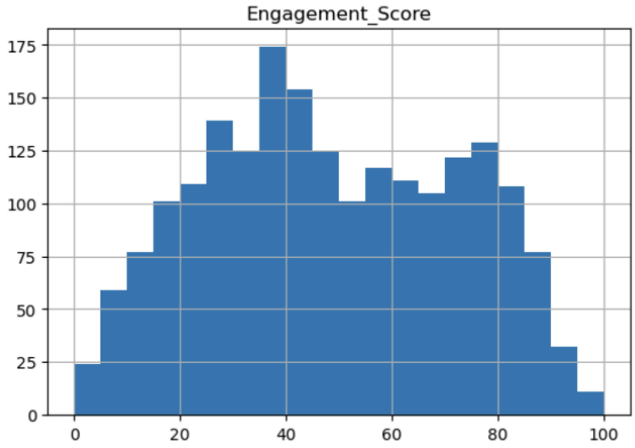
- **Demographics**:
    - Age: Customers primarily range from **20-40 years**.





    - Gender: Slightly higher percentage of **female customers**.
    - Geography: Predominantly from **urban** and **suburban** areas.
- **Spending Behavior**:

- Spend distribution is **bimodal**, with clusters of **low spenders (< $30)** and **high spenders (> $100)**.



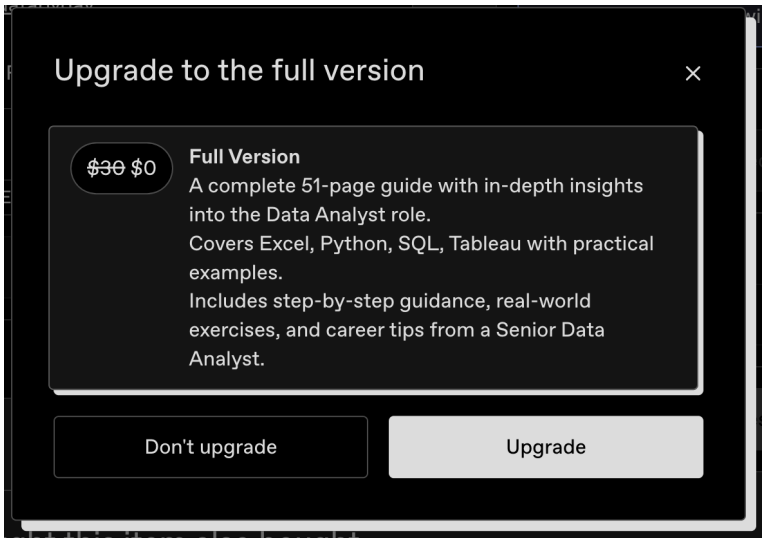
- Engagement scores have a **right-tailed distribution**, indicating a majority of users exhibit low engagement, with fewer highly engaged users.

- **Top Products**:

- The **eBook** is the most purchased product.
- The **data course** follows as the second most popular offering.

Upsell A/B Test Analysis

A/B test was conducted on users to determine the effectiveness of upsell messaging when adding the *How To Become A Data Analyst* eBook sample to their cart. The treatment group receives the upsell message below while the control does not receive the message when selecting the trial version.



Key Results:

Group	Conversion Rate
Control	28.2%
Treatment	59.6%

- **P-value**: **< 0.0001**, indicating a statistically significant difference in conversion rates.

Recommendations:

1. **Deploy Upsell Messages**:
    - Roll out the upsell message to all users adding the eBook sample to their cart.
2. **Subgroup Analysis**:
    - Identify demographic or behavioral segments where upselling is most effective. We may also want to see if there is a subgroup, such as younger demographics in which the upsell message did not make any changes. For example, we may need other strategies outside of upsell messages to convince this demographic to make purchases.

Key Benefit: Improved conversions and revenue through targeted upsell strategies.

Funnel Analysis

Stage Breakdown:

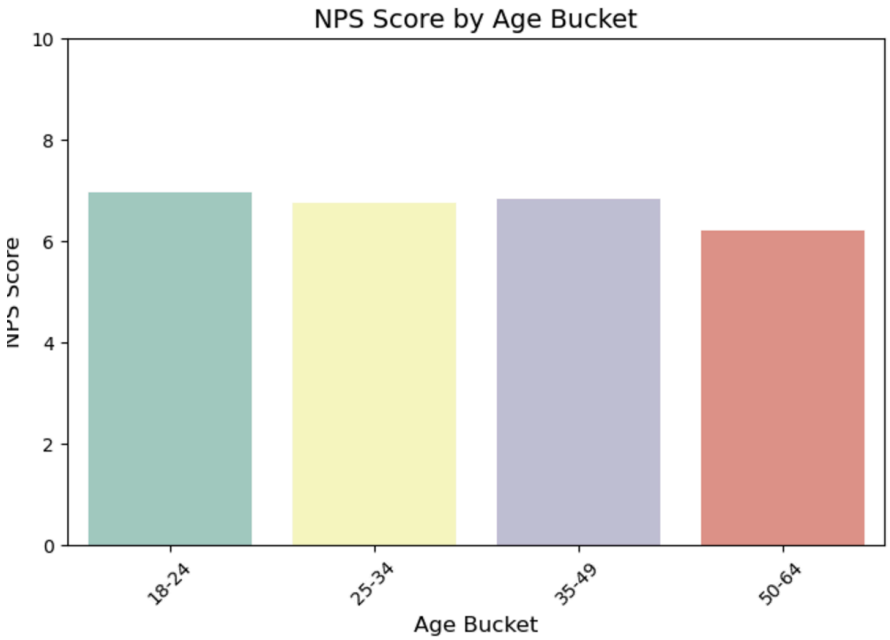
Stage	Count	Percentage	Conversion Rate
Lead	918	45.9%	-
Engaged	633	31.65%	68.95%
Converted	449	22.45%	70.93%

Insights:

1. **Strengths**:
    - High **Lead → Engaged** (68.95%) and **Engaged → Converted** (70.93%) conversion rates indicate effective top- and mid-funnel strategies.
2. **Opportunities**:
    - Investigate the **31% of leads** and **29% of engaged users** who do not convert:
        - Use segmentation to identify barriers (e.g., price sensitivity, unclear benefits of products).
        - Potentially test whether current price points deter leads from purchasing a product outside of the initial sample.

Product Performance

- **Average Spend Per User**: $48.14.
- **Distribution of Spend:** Most people spend $30 or less.



- **Net Promoter Score (NPS)**:
    - Scores across products are consistent (~7), with opportunities for improvement, especially for the 50-64 group. However, this group is not the target audience as databyday was initially aiming to support those from 18-30.

Recommendations:

1. Understand product purchases in the lower NPS individuals to improve the overall average NPS.
2. Enhance high-impact product features based on user feedback. Supplement the quantitative NPS measure with qualitative comments from customers could be helpful in understanding how databyday could improve products and services.

Predictive Modeling for Retention

After comparing multiple models, including a dummy classifier, Random Forest, and XGBoost, the recommendation is based on performance metrics, model complexity, and computational efficiency.

Model Development

First, I applied one-hot encoding to the location tier, which was originally a categorical variable. Next, I created a correlation matrix to evaluate the relationships between the features and the target variable, "Converted", identifying both strong and weak correlations. To address potential

multicollinearity among the features, I calculated the Variance Inflation Factor (VIF). Variables with VIF values greater than 10, indicating high multicollinearity, were removed. For example, the engagement score was excluded due to its high VIF, signaling strong collinearity with other features. This process ensured a more robust and interpretable model.

Model 0: Dummy Classifier (Baseline)

- **ROC AUC**: 0.50
- **Accuracy**: 0.59
- **Key Observations**:
    - The dummy classifier serves as a naive baseline.
    - It demonstrates how majority-class predictions would perform on the dataset.

Model 1: Random Forest (Initial)

- **ROC AUC**: 0.63
- **Accuracy**: 0.68
- **Key Observations**:
    - Random Forest showed an improvement over the dummy classifier, indicating that the model captures more relationships in the data.

Model 2: Random Forest (Additional Features)

- **ROC AUC**: 0.81
- **Accuracy**: 0.74
- **Key Observations**:
    - Random Forest with additional features showed a large improvement over the dummy classifier and the initial model, indicating that the model captures more relationships in the data.

Model 3: Random Forest (Recursive Feature Elimination)

- **ROC AUC**: 0.81
- **Accuracy**: 0.75
- **Key Observations**:
    - Recursive Feature Elimination (RFE) and hyperparameter tuning further refined the model, improving both recall and precision for the minority class (positive predictions).

K-Fold Cross Validation

- **Best CV ROC AUC**: 0.80
- **Test ROC AUC**: 0.815
- **Accuracy**: 0.75
- **Key Observations**:
    - The cross-validated model confirmed the robustness of the results, indicating that the model performs well across different data splits.

Model 4: XGBoost

- **ROC AUC**: 0.81
- **Accuracy**: 0.74
- **Key Observations**:
    - XGBoost matched Random Forest's performance in terms of ROC AUC and accuracy.
    - As a boosting method, it is more computationally intensive but handles complex relationships in the data better, especially in cases where additional optimization may be required.

Recommendation: Random Forest (Recursive Feature Elimination)

Why Choose Random Forest?

1. **Similar Performance**:
    - Both Random Forest and XGBoost achieved comparable performance, with ROC AUC scores around **0.81** and accuracy at **0.74-0.75**.
    - The difference in performance is negligible.
2. **Computational Efficiency**:
    - Random Forest is faster to train and tune, especially with larger datasets or multiple hyperparameter candidates.
3. **Ease of Use**:
    - Random Forest has fewer hyperparameters to tune and is easier to interpret, especially when using feature importance for explainability.

Next Steps

- Proceed with the **Random Forest model (Recursive Feature Elimination)**, as it balances performance and efficiency effectively.
- Utilize the selected features from RFE for streamlined predictions.
- If future datasets grow in complexity, consider revisiting XGBoost for potential gains in capturing non-linear relationships.

Action Plan

Immediate Steps:

1. **Upsell Messaging**:
    - Roll out proven upsell strategies across the user base.
2. **Funnel Optimization**:
    - Address key drop-off points such as when people select a sample ebook, but do not go on to purchase the full ebook.

Medium-Term Goals:

1. **Retention**:
    - Use predictive models to identify at-risk users and intervene early, such as with reminder marketing emails or discounts.
2. **Product Development**:
    - Leverage insights from NPS and perhaps gather qualitative feedback to understand how to enhance offerings.

Long-Term Goals:

1. **Scalable Marketing Strategies**:
    - Focus on high-performing segments and refine campaigns for maximum ROI.
2. **Continuous Improvement**:
    - Regularly update models and strategies to align with evolving business goals.