Customer Analytics Final Report
Dataset: IBM Watson Marketing

Team Name: SwitchMaster 5000
Team Members: Aishwarya Panse, Farin Fukunaga, Jann Ang, Maggie Ding, Queenie Chao

EXECUTIVE SUMMARY

This project analyzes the IBM Watson Marketing Customer Value Dataset, focusing on car insurance dynamics and customers with expiring policies. The dataset has 9,134 rows and includes each customer's insurance policies, claims, and marketing interactions.

Two key models were employed to derive insights from the dataset. First, a Customer Lifetime Value (CLV) prediction model utilized a semi-log approach with variables such as employment status, policy details, and renew offers. It demonstrated effectiveness within the mid-range but tended to underestimate in extreme values. Insights revealed the significance of tailoring strategies for employed customers, focusing on those with a higher policy count, and refining renewal offers.

Second, a response rate prediction model, utilizing logistic regression, identified demographic and sales-related factors influencing customer response rates. Retired and highly educated customers responded more often, while certain offers and sales channels yielded lower response rates. Policy details were not predictive, emphasizing the role of demographics and sales channels in shaping response rates. Strategic recommendations include refining renewal offers and investing in specific sales channels and offers for optimal customer engagement.

In summary, this assignment analyzes the car insurance landscape using the IBM Watson Marketing Customer Value Dataset It presents a strategic roadmap for optimizing policy renewals and customer value.

BUSINESS PROBLEM

1. **New Customer Acquisition**

   Which customers are the most valuable (CLV) and how do we identify prospects that will be valuable customers?

   The aim is to identify the most valuable customers in terms of Customer Lifetime Value (CLV). The objective is to develop strategies for new customer acquisition by identifying and targeting individuals with high potential CLV. We will analyze customer demographics, behaviors, and other relevant factors.

2. **Renewal and Churn Prevention**

   How do we drive policy renewals? What renewal marketing efforts are working?

   The goal is to improve the rate of policy renewals and identify effective marketing efforts. This will involve analysis of customer satisfaction, customer experience, marketing campaigns, incentives, and other communication channels.

METHODOLOGY

DATA CLEANING & PREPROCESSING

- All 9134 entries were examined for null values and missing data (see Figure 2 in Appendix).
- Categorical variables such as Gender and Marital Status were converted into dummy variables.

MODEL SELECTION & VALIDATION

- Semi-log and Logistic Regression: These models were chosen for their efficacy in handling specific aspects of the dataset. Semi-log regression is apt for capturing non-linear relationships, while logistic regression is tailored for binary outcomes.

ANALYTICAL TOOLS & VISUALIZATION TECHNIQUES

- R Studio's packages such as ‘glm’, ‘dplyr’ and ‘ggplot2’ were used in our regression analysis.
- Data manipulation and visualization were streamlined using Anaconda, with packages like pandas, NumPy, seaborn, and matplotlib enabling efficient data handling and the creation of visualizations.
- For initial data exploration, we used histograms & boxplot to understand distributions, Q-Q plots for normality checks, and scatter plots to discern variable relationships for instance.
- Interactive visualizations were crafted with Python's Bokeh and Plotly libraries.

TRAINING & TESTING DATASET

The dataset was split into random training (80%) and testing (20%) groups in excel. The final split was selected due to the two groups having relatively even mean CLV and mean response rates.

Group	Name	n	%	sum_clv	sum_response	mean_clv	mean_response
0	Training	7321	0.80	58652614	1049	8011.56	0.1433
1	Testing	1813	0.20	14464513	259	7978.22	0.1429

DATA SUMMARY & VISUALIZATIONS

The dataset comprises 24 columns, including 2 metrics designed for customer performance evaluation and 22 variables associated with customer data.

DEPENDENT VARIABLES

The 2 metrics serve as the dependent variables within the predictive models are as follows:

- **CLV (Customer Lifetime Value):** CLV means the total revenue the car insurance company can expect to bring in from the customer over their lifetime. It’s the dependent variable in the predictive model for identifying the most profitable customers for the company.

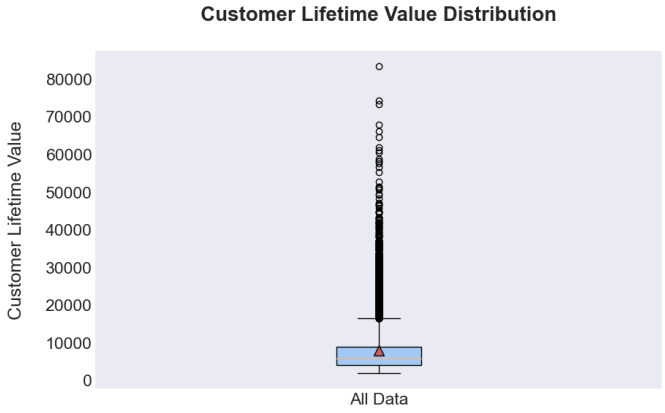
- **Response:** The *Response* column records whether the customer responded to the policy renewal marketing call. This categorical variable is converted into 2 dummy variables: "*1*" represents "*Yes, did respond"*, and "*0*" represents "*No, didn't respond*". It is the dependent variable that indicates whether a customer responded to a marketing campaign.

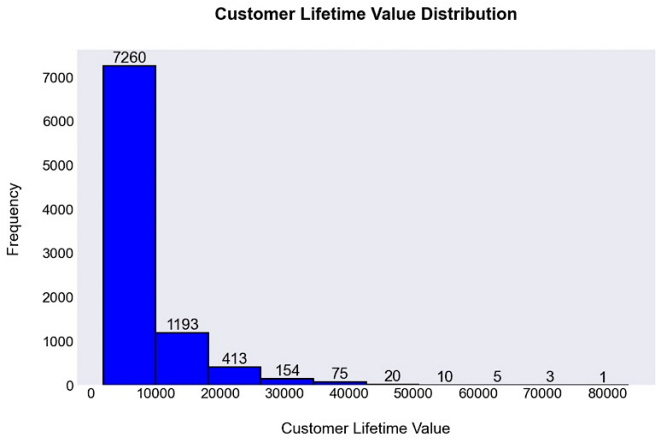
PREDICTED (INDEPENDENT) VARIABLES

The 22 predictive independent variables have been segmented into customers' demographics, auto insurance status, and policy records. They reflect individuals' behavior patterns and are highly probable to affect their insurance premium, and consequently, have correlations with their value to the company and their inclination to buy certain auto insurance products. These variables are either numerical or categorical, which makes it possible to include them in regression models. They are:

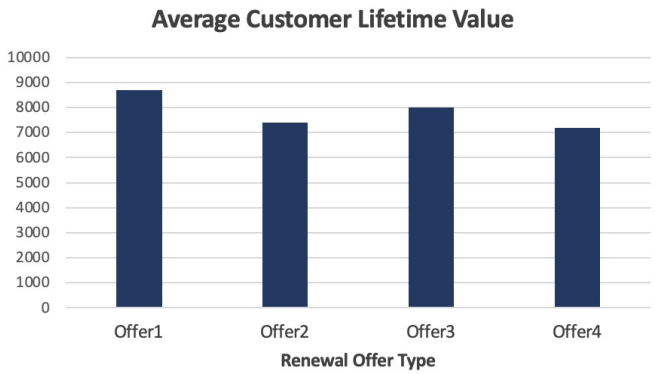
- **Demographics & Basic Information:** Customer ID (Car Plate Number), State (Living State), Education (Education Level), Employment Status, Gender, Income, Location Code (Type of Living Area), Marital Status
- **Auto Insurance Status:** Coverage (Auto Insurance Coverage Level), Effective To Date, Monthly Premium Auto, Number of Open Complaints, Renew Offer Type, Sales Channel, Vehicle Class, Vehicle Size
- **Policy & Claim Records:** Months Since Last Claim (How Long Since Last Claim), Months Since Policy Inception (Duration Since Policy Inception), Number of Policies, Policy Type (Purpose of Vehicle Use), Policy (Policy Offer Under Different Policy Types), Total Claim Amount

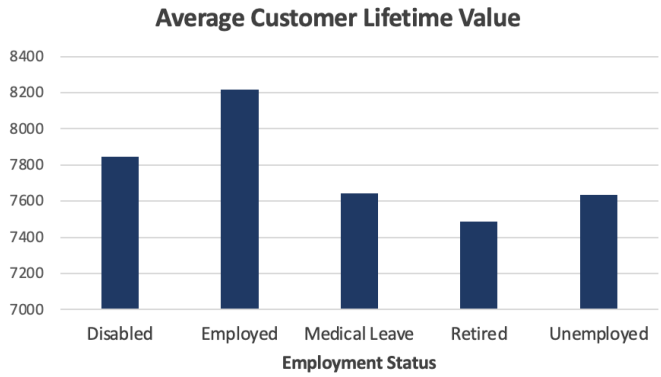
CLV DISTRIBUTION DESCRIPTIVE STATISTICS



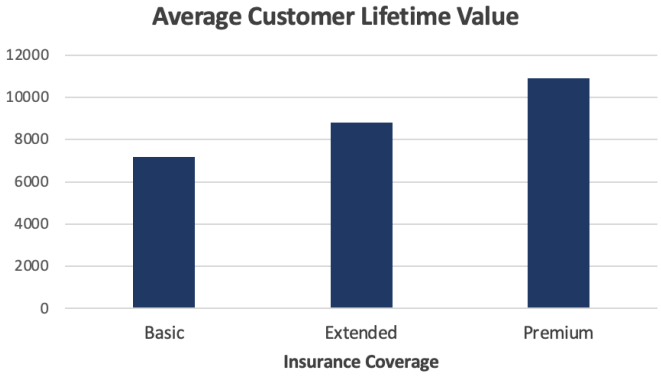


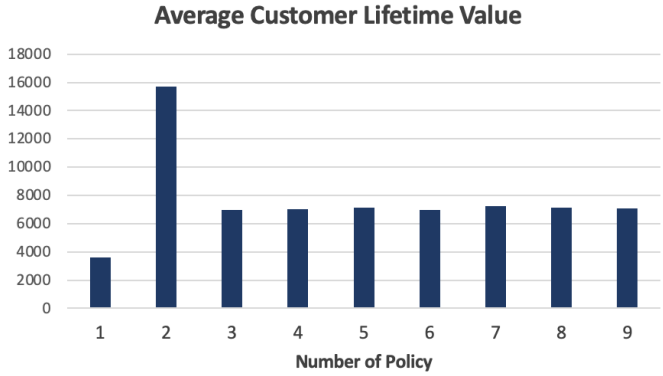
CLV Distribution: The median CLV is around $7,000, while the mean CLV is approximately $10,000, indicating a right-skewed distribution. This skewness suggests a significant portion of customers have a CLV below the average, with over 80% of customers having a CLV under $10,000.





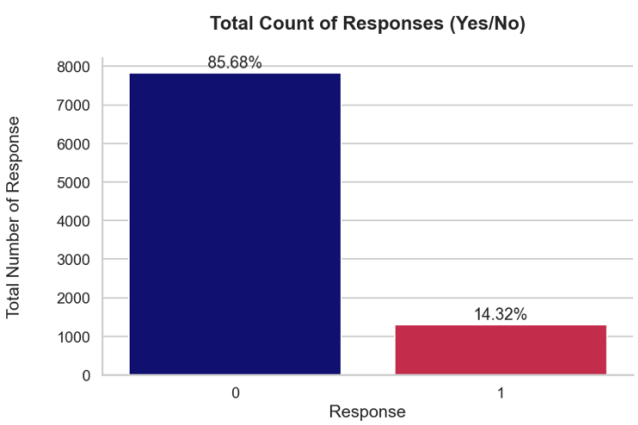
- **CLV vs. Renewal Offer Type:** Offer1 exhibits the highest average CLV, which suggests that the terms of Offer1 are slightly more aligned with the needs or preferences of the most valuable customers. The minimal variation in average CLV among the different offer types suggests that customers are not targeted based on CLV.
- **CLV vs. Employment Status:** The evident variance in CLV across employment statuses underscores the importance of employment as a segmentation variable in customer value analysis as it stands out as the highest among the categories, suggesting that employment status is a strong indicator of customer value in this context.

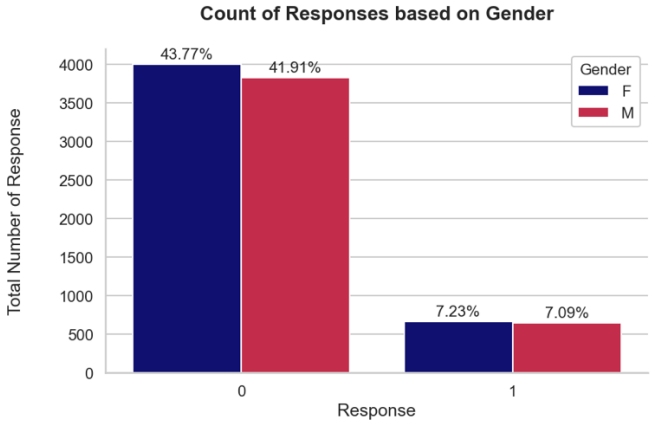




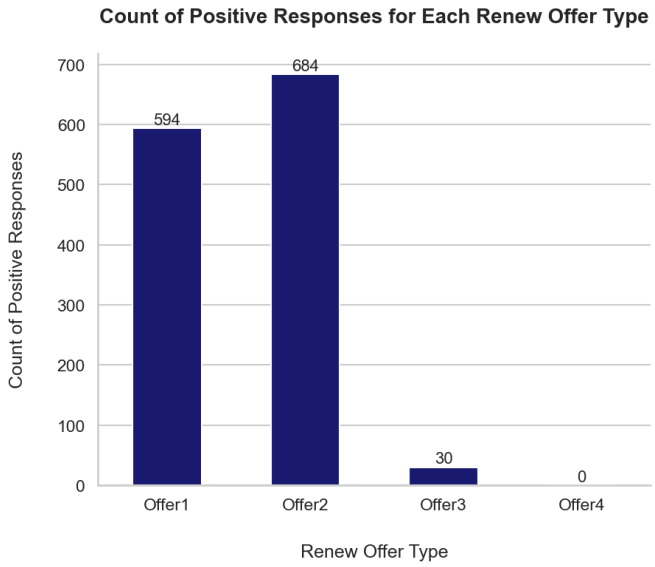
- **CLV vs. Insurance Coverage:** A clear positive correlation is visible between the comprehensiveness of insurance coverage and CLV which suggests that customers willing to invest in more comprehensive (Premium) coverage may be more valuable in the long term. Given the high CLV of Premium coverage clients, enhancing the features and services of this tier could increase its attractiveness and adoption.
- **CLV vs. Number of Policy:** The anomalous spike in CLV for customers with two policies may indicate a sweet spot where customers find optimal value or a package offering that resonates well with their needs. Customers with three or more policies tend to be corporate customers who have moderate CLVs compared to individual customers.

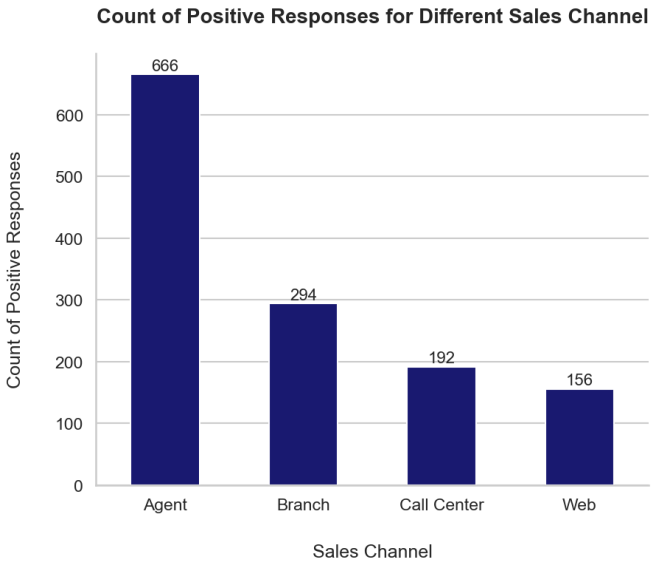
RESPONSE RATE TO MARKETING CALL



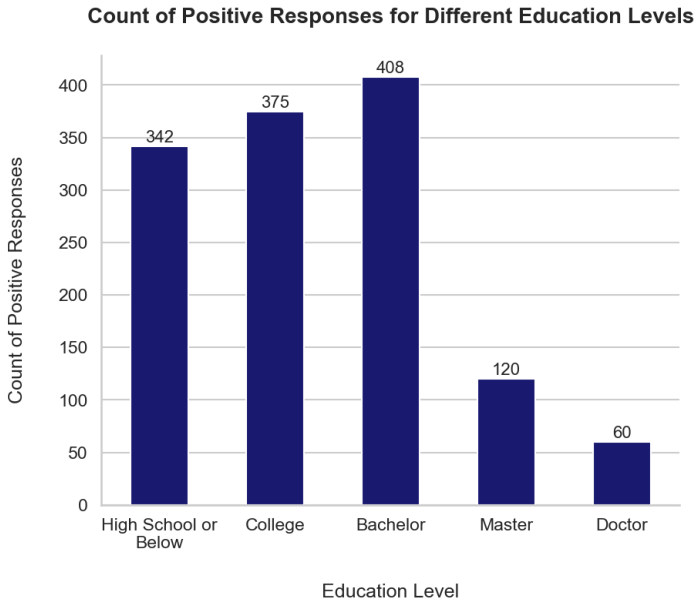


- **Response Rate (Yes/No):** Analysis of marketing responses indicates that while a substantial majority (85.68%) of customers did not respond to marketing calls, 14.32% did engage. This response disparity offers an opportunity to refine targeting strategies to enhance engagement.
- **Response Rate vs. Gender:** Males and females will respond to a marketing call in roughly equal ratios.

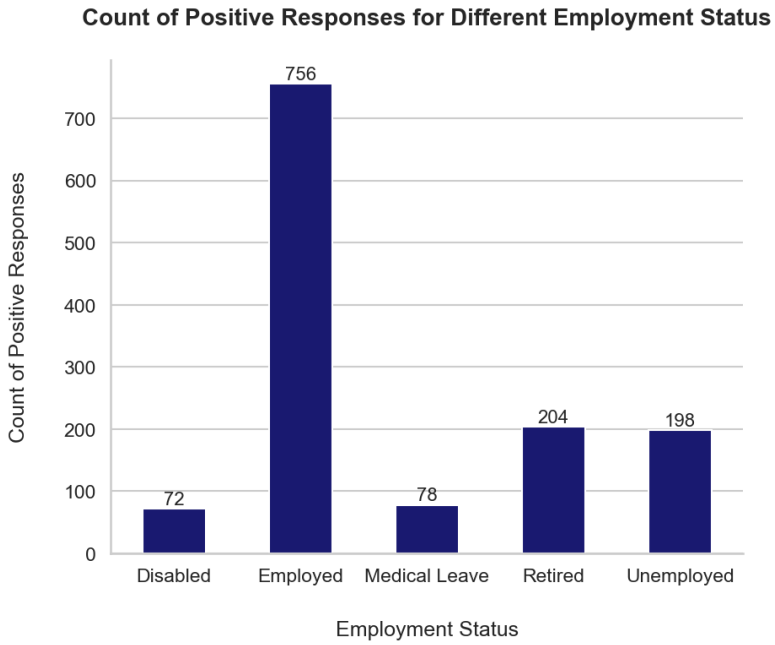




- **Response Rate vs. Renew Offer Type:** Customers have answered marketing calls for offer 1 and 2 but offer 3 has significantly fewer positive responses at 30 and offer 4 received no positive responses at all.
- **Response rate vs. Sales Channel:** Sales channel through agent received the highest number of positive responses with 666 while online channel received the fewest positive responses with 156.



- **Response Rate vs. Education Level:** The highest number of positive responses is at the Bachelor level, followed closely by College. The number of positive responses decreases significantly at the Master's level and even more so at the Doctoral level. This could suggest that individuals with advanced degrees are less responsive to the marketing strategies employed or that they represent a smaller segment of the customer base.



- **Response Rate vs. Employment Status**: The tallest bar is for "Employed" indicating it has the highest count of positive responses, while the shortest bars are for "Disabled" and "Medical Leave", indicating the lowest counts of positive responses. The bars for "Retired" and "Unemployed" are of similar height, indicating similar counts of positive responses for these categories.

MODELING #1: PREDICTING CUSTOMER LIFETIME VALUE (CLV)

THE MODEL

Customer Lifetime Value follows a skewed distribution, with a few customers having very high values. Employing semi-log transformation helps normalize the data, making it suitable for linear regression modeling. The semi-log transformation increased r-squared by 7 percentage points compared to a similar linear model. Below is a summary of the final model:

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.0757 -0.4592 -0.1114  0.2659  1.8509

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                7.9352335  0.0254911 311.295  < 2e-16 ***
Coverage_Ext               0.0431497  0.0154534   2.792  0.00525 **
Coverage_Pre               0.0239962  0.0259835   0.924  0.35577
Employ_Unemp              -0.1195313  0.0158453  -7.544 5.12e-14 ***
Employ_Dis                -0.0668636  0.0329290  -2.031  0.04234 *
Employ_Med                -0.0753617  0.0317513  -2.373  0.01765 *
Employ_Ret                -0.1183461  0.0388741  -3.044  0.00234 **
Monthly.Premium.Auto       0.0081876  0.0002166  37.801  < 2e-16 ***
Number.of.Open.Complaints -0.0286805  0.0072762  -3.942 8.17e-05 ***
Number.of.Policies         0.0522081  0.0027538  18.959  < 2e-16 ***
Policy_Cor                -0.0241376  0.0164113  -1.471  0.14139
Policy_Spe                 0.0912785  0.0318373   2.867  0.00416 **
Renew_02                  -0.1299291  0.0160120  -8.114 5.68e-16 ***
Renew_03                  -0.0679791  0.0198515  -3.424  0.00062 ***
Renew_04                  -0.1471387  0.0224736  -6.547 6.26e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5647 on 7306 degrees of freedom
Multiple R-squared:  0.2574,    Adjusted R-squared:  0.256
F-statistic: 180.9 on 14 and 7306 DF,  p-value: < 2.2e-16
```

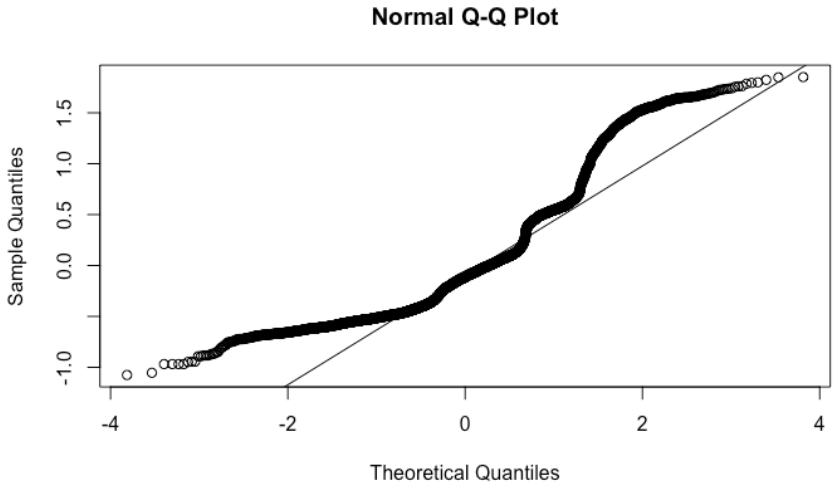
The dependent variable, ‘CLV’, was numeric, indicating the total revenue potential from a customer over time.

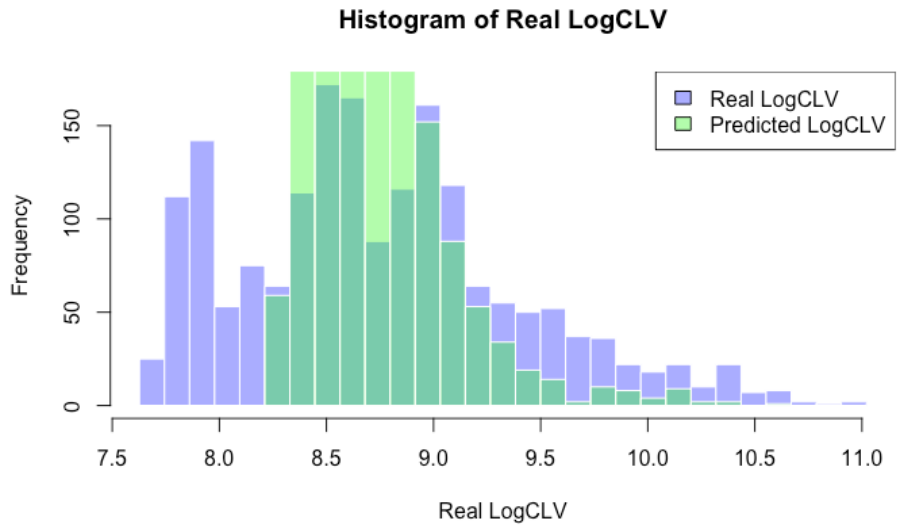
The independent variables used provides a multi-faceted view of the factors influencing customer decisions. They are:

- ‘Coverage’ (Default: Basic)
- ‘Employment Status’ (Default: Employed)
- ‘Monthly Premium Auto’
- ‘Number of Open Complaints’
- ‘Number of Policies’
- ‘Policy’ (Default: Personal)
- ‘Renew Offer Type’ (Default: Offer1)

DIAGNOSIS

Although our model effectively predicts CLV within the mid-range, we've observed a tendency to underestimate in extreme values. This discrepancy becomes apparent when examining the normal QQ plot and the histogram provided below.





KEY FINDINGS

- **Employment Status:** Employment status emerges as a substantial driver of Customer Lifetime Value (CLV), showcasing an increase when customers are employed. This underscores the importance of tailoring strategies to capitalize on the higher CLV associated with employed customers.
- **Number of Policies:** The number of policies exhibits a positive impact on CLV, as indicated by the positive coefficient. This insight suggests a strategic focus on customers with a higher policy count, as they present an opportunity for increased CLV.
- **Renew Offer:** The Renew Offer stands out as a pivotal factor influencing CLV. Coefficients for Renew_O2 to Renew_O4 are consistently negative, suggesting a potential pitfall in these offers. Hence, Renew Offer 1 positively contributes to enhancing CLV, signaling the need for refinement in how offers are matched to a customer.

MODELING #2: PREDICTING RESPONSE RATE

THE MODEL

Our approach employed logistic regression on the training dataset to model response rates, a method chosen for its efficacy in handling binary outcome variables. Below is a summary of the final model:

```
Call:
glm(formula = response_binary ~ EmploymentStatus + Renew.Offer.Type +
    Sales.Channel + Education, family = binomial(link = "logit"),
    data = data.train)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.20168    0.16581  -7.247 4.26e-13 ***
EmploymentStatusEmployed    -0.43865    0.15712  -2.792  0.00524 **
EmploymentStatusMedical Leave  0.02990    0.20624   0.145  0.88473
EmploymentStatusRetired      2.66542    0.22405  11.897  < 2e-16 ***
EmploymentStatusUnemployed  -0.91732    0.17222  -5.326 1.00e-07 ***
Renew.Offer.TypeOffer2       0.67095    0.07867   8.528  < 2e-16 ***
Renew.Offer.TypeOffer3      -2.10124    0.21631  -9.714  < 2e-16 ***
Renew.Offer.TypeOffer4     -16.78815  222.92553  -0.075  0.93997
Sales.ChannelBranch         -0.66737    0.09390  -7.107 1.19e-12 ***
Sales.ChannelCall Center    -0.48808    0.10430  -4.679 2.88e-06 ***
Sales.ChannelWeb            -0.55519    0.11843  -4.688 2.76e-06 ***
EducationCollege             0.16489    0.09492   1.737  0.08237 .
EducationDoctor              0.48699    0.18139   2.685  0.00726 **
EducationHigh School or Below  -0.04270    0.09850  -0.434  0.66465
EducationMaster              0.23962    0.14446   1.659  0.09716 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6016.2  on 7320  degrees of freedom
Residual deviance: 4850.9  on 7306  degrees of freedom
AIC: 4880.9

Number of Fisher Scoring iterations: 17
```

The dependent variable, ‘Response’, was binary, indicating whether a customer would renew their policy.

The independent variables used provides a multi-faceted view of the factors influencing customer decisions. They are:

- ‘Employment Status’ (Default: Disabled)
- ‘Offer Type’ (Default: Offer 1)
- ‘Sales Channel’ (Default: Agent)
- ‘Education’ (Default: Bachelors)

Other variables were considered, but they were not included in the final model due to lack of statistical significance, explanatory power, or relevance to the problem.

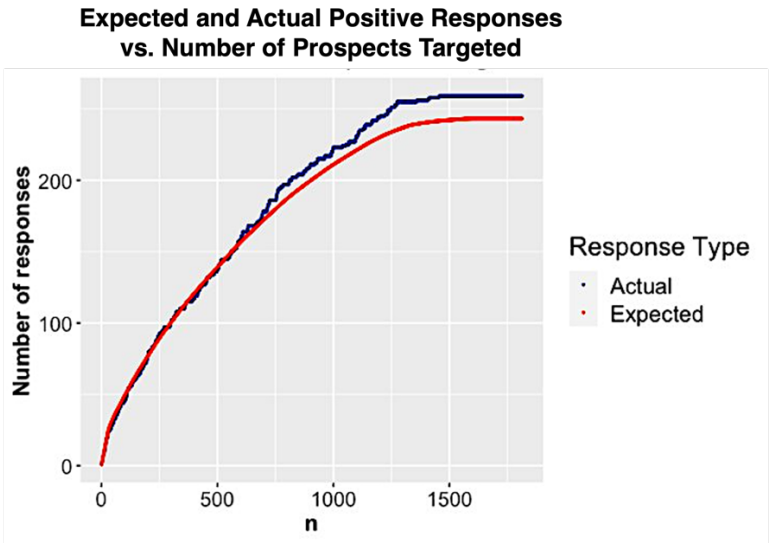
DIAGNOSIS

Below is the confusion matrix of predicted and actual results on the testing dataset:

	Actual		
Predict	0	1	Row Total
0	1546	234	1780
1	8	25	33
Column Total	1554	259	1813

This model did a good job of predicting negative responses but struggled with false negatives (only 25 of 259 positive responses were correctly predicted). Adding additional independent variables from the dataset did not improve the false negative rate. The likely cause of this is omitted variable bias - there were factors that influenced response rates that were not captured in this dataset.

Below is the expected response curve vs the actual number of responses for the testing dataset:



As shown above, the model did a fairly good job of predicting response from the ~500 customers who were mostly likely to respond. However, the model underestimates response rates from the remaining portion of the dataset. This is likely due to fewer strong signals such as employment status being “retired” or education level being “doctorate” in this customer group.

KEY FINDINGS

The significant variables for this model were either demographic (employment status and education) or sales related (offer type and sales channel). In general, retired and highly educated (doctors or master’s degree) customers responded more often. It is an open question whether being retired and having an advanced degree makes a customer more receptive to insurance marketing or whether these factors are proxies for age, and older people respond more often. For sales related factors, offers 3 and 4 as well as the branch, call center, and web channels each lead to lower response rates than their alternatives. Therefore, we’d recommend investing more into offer 1 and 2 as well as agent marketing. Interestingly, policy details such as policy type, number of policies, number of claims and CLV were not accurate predictors for response rate. This suggests that response rate is a function of the customer’s demographics and offer/sales channel rather than a function of what insurance products the customer uses.

RECOMMENDATIONS

Based on the regression analysis results and key findings presented above, here are some specific recommendations to improve marketing decisions:

OPPORTUNITY AREA #1: INCREASE CLV

1. **Enhance Focus on Employment Status:**

   Given that employment status is a significant contributor to CLV, marketing initiatives should be customized based on employment categories. For example, create tailored communications for employed individuals emphasizing the security a policy provides against potential income loss from an auto accident, or partner with larger employers to target their workers.

2. **Leverage Policy Count:**

   The positive impact of the number of policies on CLV suggests that customers with multiple policies are more valuable. Marketing strategies could include bundle offers or discounts for customers who hold or add multiple policies can incentivize them to increase their business with the company leading to higher retention rates and increase CLV. It capitalizes on the convenience factor and loyalty—customers are likely to appreciate the simplicity of managing all their insurance needs through one provider. This approach can also lead to economies of scope, where the cost of serving a customer with multiple policies is less than the combined cost of serving several customers with single policies.

3. **Streamline Policy Upgrades:**

   The company should make it easy for customers to upgrade or add policies online or via their agent. These pathways should be clearly documented and communicated through all marketing channels. Along with excellent customer service, this will lead to increased customer satisfaction and higher CLV.

OPPORTUNITY AREA #2: IMPROVE RESPONSE RATES

1. **Targeting elderly retired and highly educated people through their personal agent**

   Capitalizing on personal touch that can be very effective with older demographics, who may value relationships and direct interaction more than digital channels. Additionally, higher education often correlates with a greater ability to understand complex information. By routing marketing messaging through personal agents, we can ensure that customer receive the information they need to make informed renewal decisions.

2. **Reevaluate sales channels like branch, call center and web and refine renewal offers 3 and 4**

   Gather sentiment analysis on why these offers are not performing well provides data-driven insights that can be used to restructure offers. For instance, it might reveal that customers feel their needs are not being addressed adequately in call centers, or that the web interface is not user-friendly. These changes could be A/B tested with existing offers to measure if changes can lead to improved outcomes. Further experiment with custom offers based on customer profiles may increase the likelihood of renewal.

FUTURE ANALYSIS

For future analysis and implementation to refine marketing strategies and improve customer retention, consider the following detailed approaches:

1. **Longitudinal Tracking:**

   Implementation: Establishing a longitudinal tracking system involves creating a customer data platform that captures and integrates all customer interactions, transactions, and engagements over time across various channels.

   Analysis: Compare initial predictions with actual customer behavior. Were there any renewals from individuals who did not react in this dataset? Is there any evidence that the campaigns were successful? This includes whether non-responsive customers in the dataset eventually renewed and if there was a real lift in response rates from campaigns.

   Action: Based on these insights, marketing initiatives can be realigned to focus on the most effective channels and touchpoints. For instance, if it is observed that customers frequently lapse after a certain policy

period, preemptive engagement strategies could be implemented just before this critical point to improve retention.

2. **Clustering and Micro-Segmentation:**

   Implementation: Utilizing advanced analytics and machine learning algorithms, such as k-means clustering or hierarchical clustering, the company can segment the customer base into smaller, more homogeneous groups based on various attributes, like behavior patterns, policy preferences, and demographic details.

   Analysis: Once clusters are formed, the company can analyze the unique characteristics and preferences of each group. For instance, one cluster may include young professionals who prefer digital interactions and are interested in auto insurance, while another may consist of retirees more responsive to health insurance offers via direct mail.

   Action: Develop targeted marketing strategies for each cluster to optimize the effectiveness of outreach efforts. For example, a campaign for young professionals might highlight the convenience of mobile app services, while a campaign for retirees might emphasize stability and customer service.

3. **Loyal Program Development:**

   Implementation: Implement a loyalty program that rewards customers for various interactions and milestones, such as renewing policies, referring new customers, or purchasing additional products.

   Analysis: Segment customers based on their engagement level with the loyalty program and analyze the impact of different loyalty tiers on customer retention and CLV. Determine which rewards or incentives are most effective at driving desired customer behaviors.

   Action: Enhance the loyalty program based on the analysis to maximize customer retention and encourage desired behaviors such as cross-selling uptake. For instance, a high-tier loyalty member might receive free roadside assistance with their car insurance policy.

CONCLUSION

In wrapping up our deep dive into the IBM Watson Marketing analysis, it's clear we have identified pivotal factors that significantly enhance Customer Lifetime Value (CLV). Notably, employment status and the multiplicity of policies held by customers have emerged as critical drivers in augmenting CLV. Concurrently, our investigation into response rates underscores the importance of personalized engagement, particularly with the retired demographic and individuals possessing advanced educational qualifications.

The implementation of predictive modeling goes beyond its status as an innovative tool, functioning as a strategic asset in discovering clustering opportunities and enriching loyalty programs. By anticipating and solving client needs, we go beyond transactional engagements to proactive solutions, building long-term brand loyalty.

Armed with these tools and insights, the objective now is to adopt strategic initiatives that transform analytical insights into concrete financial development and nurture client relationships into long-term partnerships. The knowledge gained from this study enables us to not only respond to current market demands, but also to forecast and shape future trends. Here's to the future of informed, customer-centric marketing excellence.

APPENDIX

Figure 1: Data Understanding

```
'data.frame':   9134 obs. of  23 variables:
 $ State                    : chr  "Washington" "Arizona" "Nevada" "California" ...
 $ CustomerLifetimeValue    : num  2764 6980 12887 7646 2814 ...
 $ Response                 : chr  "No" "No" "No" "No" ...
 $ Coverage                 : chr  "Basic" "Extended" "Premium" "Basic" ...
 $ Education                : chr  "Bachelor" "Bachelor" "Bachelor" "Bachelor" ...
 $ EffectiveToDate          : chr  "2/24/11" "1/31/11" "2/19/11" "1/20/11" ...
 $ EmploymentStatus         : chr  "Employed" "Unemployed" "Employed" "Unemployed" ...
 $ Gender                   : chr  "F" "F" "F" "M" ...
 $ Income                   : int  56274 0 48767 0 43836 62902 55350 0 14072 28812 ...
 $ LocationCode             : chr  "Suburban" "Suburban" "Suburban" "Suburban" ...
 $ MaritalStatus            : chr  "Married" "Single" "Married" "Married" ...
 $ MonthlyPremiumAuto       : int  69 94 108 106 73 69 67 101 71 93 ...
 $ MonthsSinceLastClaim     : int  32 13 18 18 12 14 0 0 13 17 ...
 $ MonthsSincePolicyInception: int  5 42 38 65 44 94 13 68 3 7 ...
 $ NumberofOpenComplaints   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NumberofPolicies         : int  1 8 2 7 1 2 9 4 2 8 ...
 $ PolicyType               : chr  "Corporate Auto" "Personal Auto" "Personal Auto" "Corporate Auto" ...
 $ Policy                   : chr  "Corporate L3" "Personal L3" "Personal L3" "Corporate L2" ...
 $ RenewOfferType           : chr  "Offer1" "Offer3" "Offer1" "Offer1" ...
 $ SalesChannel             : chr  "Agent" "Agent" "Agent" "Call Center" ...
 $ TotalClaimAmount         : num  385 1131 566 530 138 ...
 $ VehicleClass             : chr  "Two-Door Car" "Four-Door Car" "Two-Door Car" "SUV" ...
 $ VehicleSize              : chr  "Medsize" "Medsize" "Medsize" "Medsize" ...
```

Figure 2: Checking Null Values in Each Column

	count_null <int>
State	0
CustomerLifetimeValue	0
Response	0
Coverage	0
Education	0
EffectiveToDate	0
EmploymentStatus	0
Gender	0
Income	0
LocationCode	0
MaritalStatus	0
MonthlyPremiumAuto	0
MonthsSinceLastClaim	0
MonthsSincePolicyInception	0
NumberofOpenComplaints	0
NumberofPolicies	0
PolicyType	0
Policy	0
RenewOfferType	0
SalesChannel	0
TotalClaimAmount	0
VehicleClass	0
VehicleSize	0

Figure 3: Heatmap of a Correlation Matrix Using Continuous Variable Only