Capstone Project

Final Submission

Sven Meydell

Introduction

Problem Statement:

Insurance premiums paid by customers are one of the main, if not only, sources of revenue for insurance companies. Losing this revenue source **could risk damaging, at best, the company's bottom line, possibly resulting in layoffs** and various cost-cutting approaches, and, **at worst, negatively affect the entire insurance industry** – both providers and reinsurers through **insufficient cash flows to offset claim payout requests**.

The goal of this analysis is therefore to **predict the probability that a customer will not renew their policy and default on their premium payment**, helping insurance agents proactively reach out to the policy holders to **follow up on premium payments** and mitigate potential future losses while maintaining a healthy cash-flow threshold.

Study Need/Purpose:

Many companies rely on recurring premium payments in order to remain profitable and continue to grow their businesses.

This becomes especially important when factoring in that most policies are substantially discounted from a margin standpoint, due to **high commissions paid to representatives in order to effectively sell the initial business**. Additional costs go into ensuring that those customers assigned a policy are deemed to be extremely reliable in making future payments while also being relatively low risk from a claims standpoint.

Obviously, a lot of costs and planning have gone into attaining each new customer, with **an expected timeframe required in order to generate a profit from the initial acquisition costs**.

If enough customers were incorrectly identified as reliable, only to later not renew their policies, the overall ramifications to the business and industry could be drastic. Since these customers sampled are already actively assigned policies, the **next best option for the business is to create early warning identification of possible defaulters and act promptly to reduce future losses and possibly save those accounts**.

Benefit to Business:

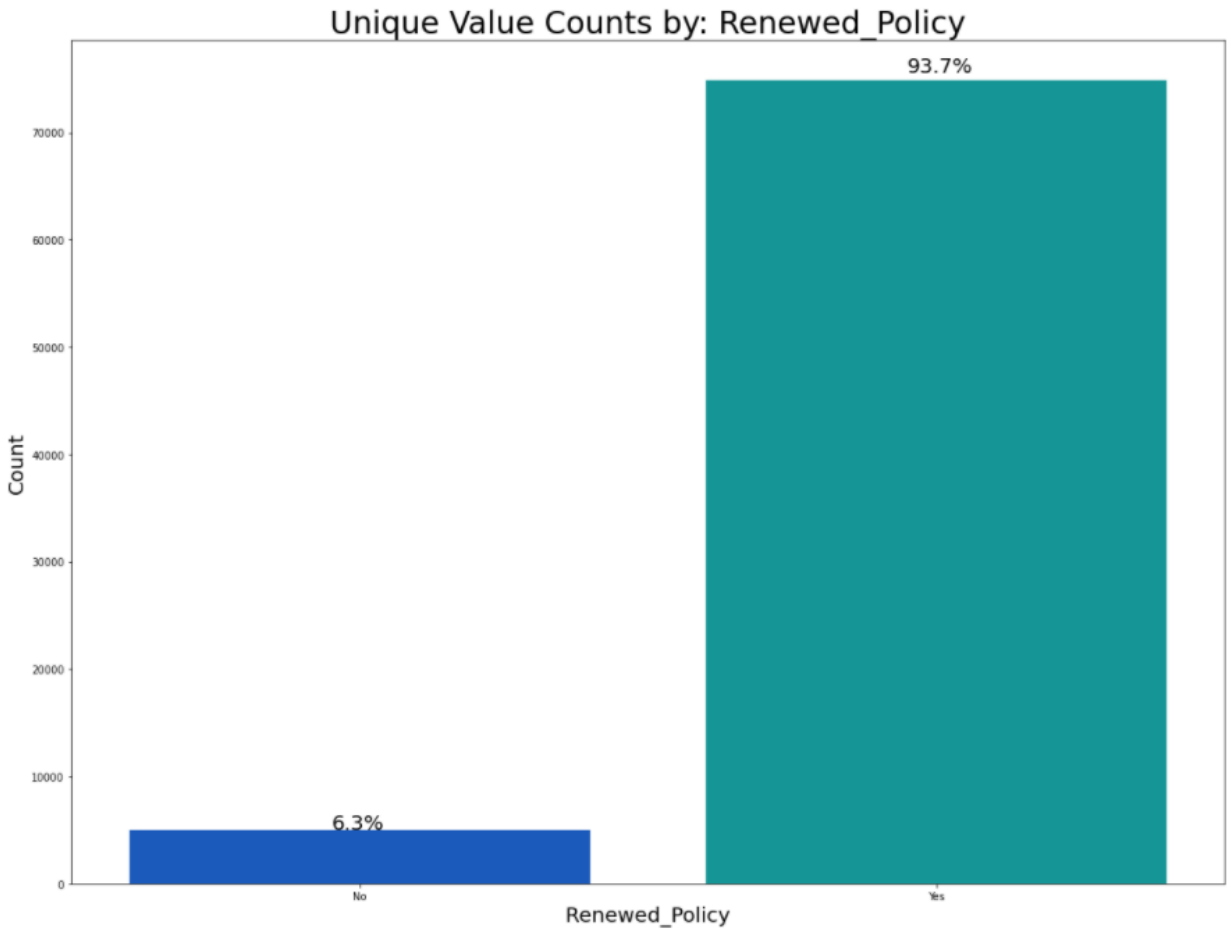
There is also the need to **secure ongoing cash flows in order to offset paying claims as needed**, or at least netting more than the amounts paid on reinsurance payments to other providers.

The entire Insurance industry is united in their reliance on consistent cash flows from reliable customers so as to keep the system of payments in/out ongoing.

From a social standpoint, any potential loss of revenue through early identification of at-risk customers stands to save the company money, **potentially saving jobs, whilst also contributing to the overall safety structure of the insurance/reinsurance partnership** that keeps everything afloat (premiums in/claims out).

Target Variable – Imbalance Class:

Although it is a positive situation in general, having almost 94% of the customers sampled renew their policies and stay active (6% not renewing), the large class imbalance makes for a **series of adjustments to any predictive modeling attempts so as to not purely pick customers who renew (94% chance of success) but rather those at risk of defaulting on their policies.** The imbalance will be addressed during the modeling building and scoring process utilizing minority scaling, class weights, and specific scoring metrics.



Data Review

Data Collection:

The dataset is a sample taken from the overall, currently active, customer database (population), was selected at random, and includes the following (cleaned/renamed) columns/variables:

- **ID**: Unique customer ID
- **Percent paid by Cash/Credit**: What % of the premium was paid by cash payments?
- **Age in Days**: Age of the customer (days)
- **Income**: Income of the customer
- **Marital Status**: Married (1)/Unmarried (0)
- **Vehicles Owned**: Number of vehicles owned (1-3)
- **Count (3-6 Months Late)**: Number of times premium was paid 3-6 months late
- **Count (6-12 Months Late)**: Number of times premium was paid 6-12 months late
- **Count (More than 12 Months Late)**: Number of times premium was paid more than 12 months late
- **Risk Score**: Risk score of customer (as it relates to likelihood of a future insurance claim)
- **Number of Dependents**: Number of dependents in the family on the customer (1-4)
- **Accommodation**: Property Rented (0)/Owned (1)
- **Number of Premiums Paid**: Number of premiums paid thus far
- **Sourcing Channel**: Channel through which customer was sourced
- **Residence Area Type**: Residence type of the customer (Rural/Urban)
- **Premium Renewal**: Variable indicating if Customer has Renewed (1) or not Renewed (0) their Policy

No specific timeframe is provided for when or how often the data was collected. Rather, it appears to be a random sample taken at a moment in time instead of a regular timeframe/frequency

Data Inspection:

Target Variable:

- **Renewal is the target column, with Non-Renewals (0) being the minority target class specifically**
- Interestingly, a combination of the 3 Count of Lates (Months) could also indicate possible at-risk customers, however the primary goal is to determine and address active customers not likely to renew their policies

Overall Data Summary:

- There are **79,853 rows** of data spread across **17 columns**
- There are no missing/null values based on initial summary analysis
- Most of the variables shown are numeric (**Int64/Float64**), with 2 showing as **Object**
    - Sourcing_Channel
    - Residence_Area_Type
- Both variables can be **converted to Categorical to save space** and allow for cleaner EDA, and later on encoded to numerical values, for modeling purposes
- There are **no Null or N/A values** within the dataset **nor any duplicated rows**

Variable/Feature Data Details:

- The id column is acting as an index with unique values for each row count - **it can kept for now and set to the training index column (for use later in uniquely identifying predicted at-risk customers)**
- The categorical columns listed above, in addition to various numerical columns have limited, discrete, counts of unique features (2 to 5), and are listed (cleaned) as follows:
    - Marital Status (2) – **Binary (Yes/No)**
    - Accommodation (2) – **Binary (Rent/Owned)**
    - Residence Area Type (2) – (Rural/ Urban)
    - Renewal (2) – **Binary (Yes/No)**
    - Vehicle Owned (3) – (1/2/3)
        - No option for 0 cars owned by customers
    - No of Dependents (4) – (1[self]/2/3/4)
    - Sourcing Channel (5) – (A,B,C,D,E)
- The **Age in Days field can be converted to years** as a specific day counter doesn't appear to offer any benefit to this analysis or the business objective
    - This should substantially lower the number of unique values to under 100 – currently 833

Statistical Insights:

- Percentage Premium Paid by Cash is a ratio spanning from 0 to 1.0, as expected
    - The distribution is Right/Positive skewed (Mean larger than Median)
    - **Of 75% of the customers sampled, only just over 50% of the premiums are paid in cash**, with credit appearing a far more popular method of payment
- Age in Days will be converted to Years, which will largely reduce the range and quantile results - maximum within 100 years or less
- Income is Right/Positive skewed (Mean larger than Median), with an **extremely large maximum income value of $90.3M**
    - This is likely a mistake and should be further examined
- All three **Count of lates (3-6 months, 6-12 months, and more than 12 months) appear largely Right/Positive skewed (Mean greater than Median)** due to the fact that most customers, usually, don't default on payment (0 count), with those few that do appearing like outliers
    - This field may offer more benefit as a single count of late payments after a certain threshold as the probability of repayment most likely deteriorates rapidly after a set amount of time
    - Further statistical analysis/testing should be performed to confirm taking this preprocessing step
- Marital Status shows a mean of roughly 50% between Unmarried (0) and Married (1)
    - Further exploratory analysis will be needed to determine whether this variable has any effect on likelihood of late payment/default
- **Vehicle Owned is normally distributed** with an equivalent Mean/Median of around 2 cars per customer (family) sampled
- Number of Dependents range from a **min of 1 (individual customer) to 4 (customer, spouse, 2 children for example)**, with customers having on average between 2 and 3 total dependents
- Accommodation shows a relatively **even split between Rented (0) and Owned (1) properties**, with a Mean of roughly 50%
- Customer Risk Scores, based on an aggregation of all aspects of a given customer's profile, indicates the (positively correlated) riskiness or likelihood of a claim, potentially increasing likelihood of renewal (having claim coverage), if customers are aware of possibly needing to claim in the future
    - There is a very small range between the lowest and highest calculated risks (Min: 91.9%, Max: 99.9%, and Median: 99.2%), **indicating that even the lower risk customers sampled carry risk scores well above 90% and are expected to claim at some point**

- Number of Premiums Paid is **relatively normally distributed**, with Mean slightly higher than Median, and a **range of 2 to 60 (annual) payments**
    - The above analysis confirmed that **both Number of Premiums and Premiums paid are annual figures**
    - An additional variable Avg. Monthly Premium will be created in the hopes of improved modelling insights and accuracy
- **Premiums Paid ($) are largely right/positive skewed (Mean larger than Median)**, primarily due to the fact that Premium calculations are subjective and based on various Customer Profile factors
    - These profiles include variables such as: Risk Score, Income, Accommodations, Dependents, etc., all of which are **run through complicated risk/reward algorithms to minimize risks while maximizing returns**
- Renewals are **slightly left/negative skewed (Median larger than Mean)** due to the fact that the **majority of the customers sampled have renewed their premiums (1)** vs. a small percentage which have yet to do so (0)
    - **The 1st, 2nd, 3rd, and 4th quantiles all indicate premium renewals (1)** throughout the entire sample distribution

Attribute Analysis:

The customer age column is confusing and offers little benefit for the analysis.
It can be **renamed to Customer Age** and adjusted divided by 365 days, for an annual equivalent.
The average calculated premiums sampled appear **too high to be monthly amounts.** It is relatively safe to assume that Premium and **Number of Premiums paid are based on annual timeframes.**

It may therefore be beneficial to **engineer an additional feature for monthly average premiums** paid by each customer sampled, which could benefit modeling performance accuracy.

It would also serve us to derive a ratio for the relationship between customer income and respective premiums paid. Since insurance companies deal all sorts of clients, especially from an Income standpoint, **it is possible that very high earning individuals are as likely to have a policy (a much larger one) as customers with far lower income** as it is sometimes heard in the media that the very wealthy and high income earners can **earn upwards of 100x or higher than their counterparts.**

It will be worthwhile to remove the Income column and instead utilize the new feature: **Income/Premium** for easier analysis and understanding of data correlations.

As mentioned above, there is **no specific Target provided**, but rather a collection of late counts (3 to 12+ months) that collectively indicate a strong likelihood of lack of repayment, or default. A new variable can be created as a **binary split of these late counts; Late (1) or Current (0).**

Changes to Feature/Variable Details (Including Data Preprocessing):

- Divide Customer Age in Days by 365 and **rename to Customer Age**
- **Renamed Initial Variables (included above)** for cleaner Cleaner/Consistent Naming
    - Primarily proper naming/spelling of feature names
- **Creating Late Account (Target) column**: setting sum of Late Monthly columns >= 0 to 0, else 1
- Dropping Premium Variable: **replaced by Avg. Monthly Premium** (Premium / 12)
- Dropping Income Variable: **replaced by Income/Premium variable**

Exploratory Data Analysis

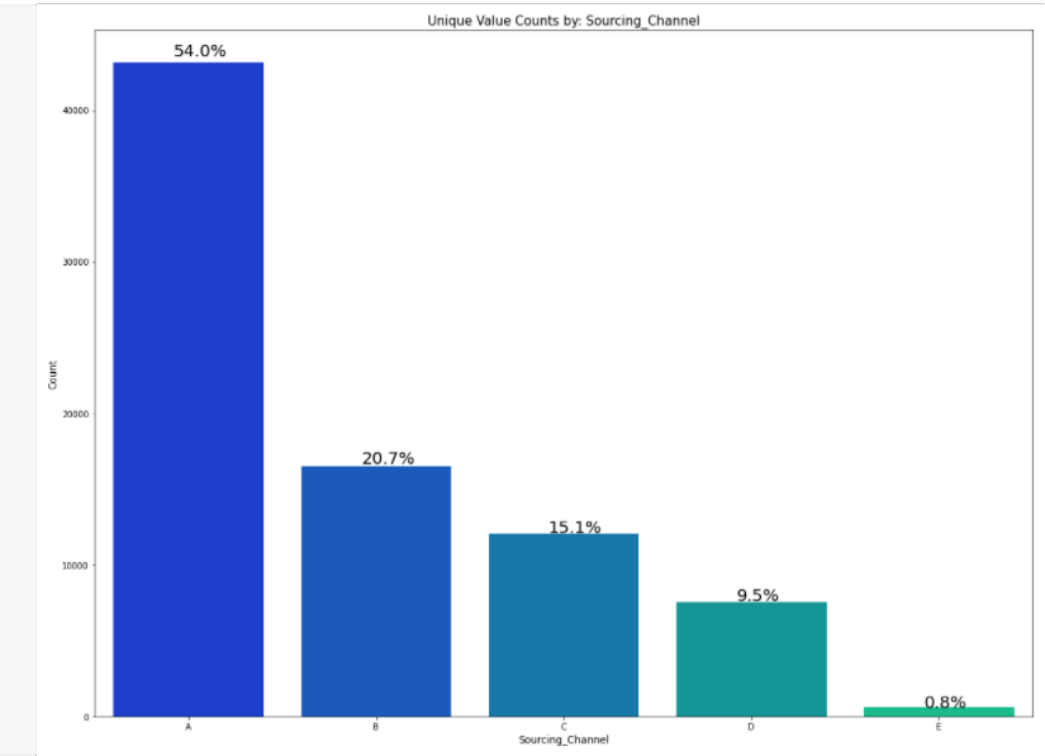
Uni-variate:

- Pandas Profiling report run for initial summary insights/analysis

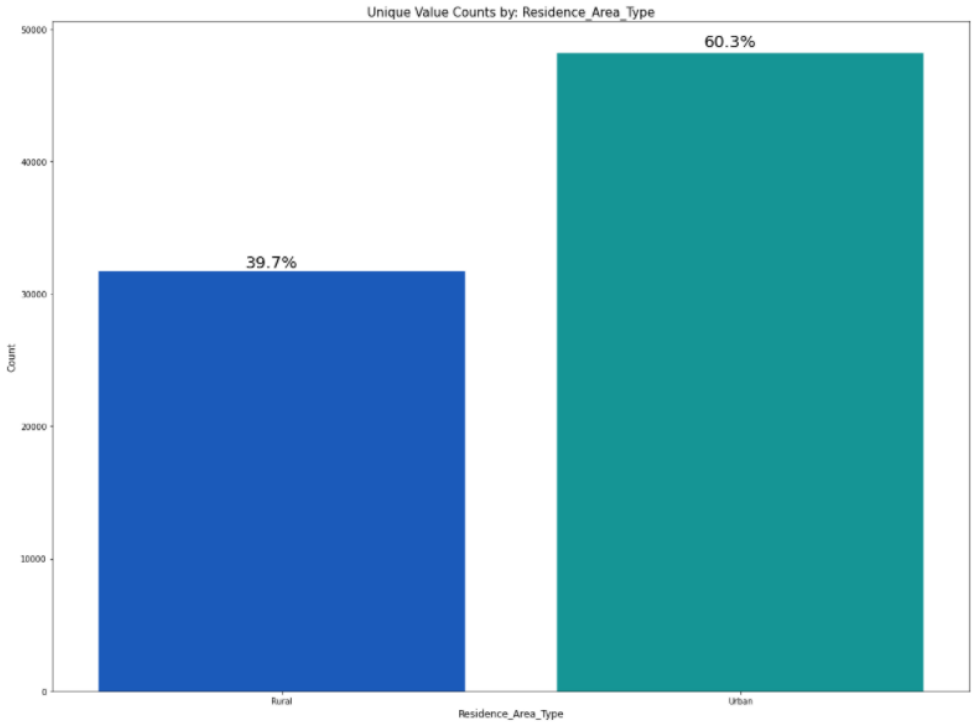
- There are **numerous categorical/discrete numerical variables** with a nearly perfect split by attribute, which is curious why the distribution is exact - this wouldn't usually be the case for a randomly sampled distribution – (approximated) examples include:
    - **Marital Status:** 50/50 split
    - **Vehicle Owned:** 33/33/33 split
    - **Number of Dependents:** 25/25/25/25 split
    - **Accommodation Rent/Owned:** 50/50 split

Categorical Visualizations:

- As mentioned already, Marital Status, Vehicle Owned, Number of Dependents, and Accommodation have relatively even splits within the data sampled
- **Sourcing Channel A is the most frequently used** for all accounts, Current and Late (44% and 9.7% respectively
    - Channels B through E **continue this trend in an ordinal fashion**

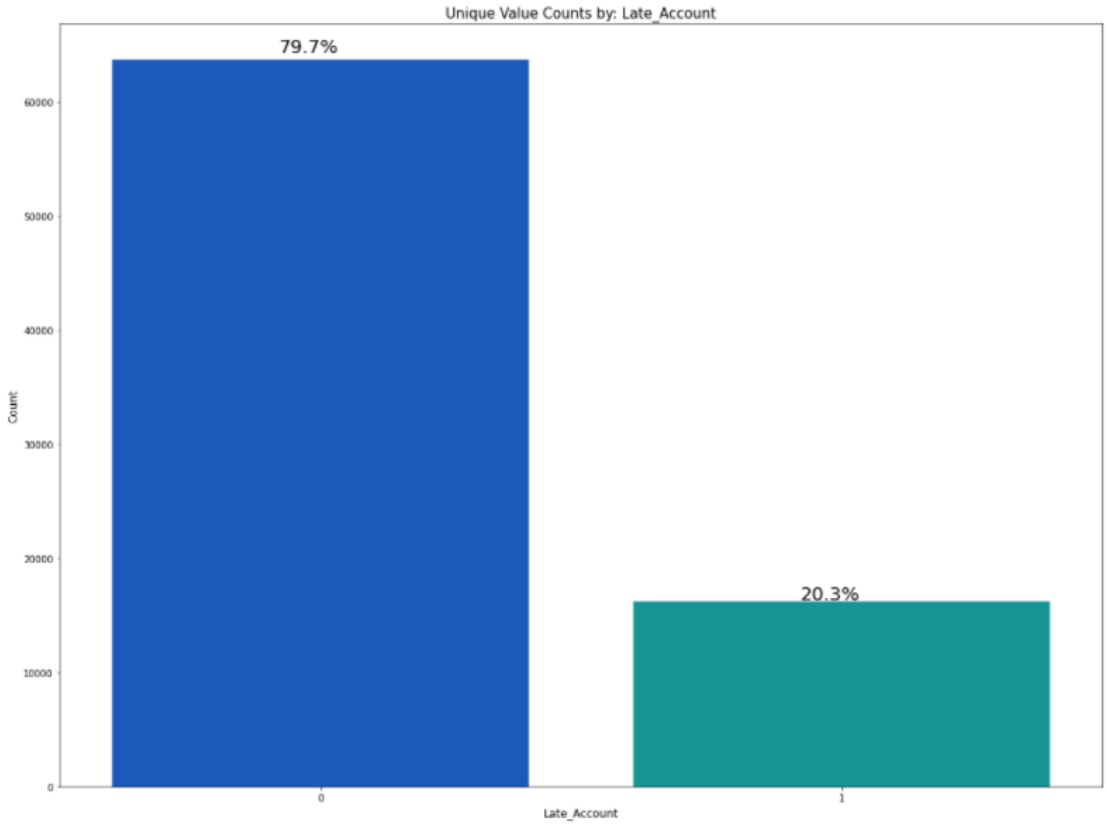


- **More of the customers sampled live in Urban areas** over Rural areas (60/40 split)



Numeric (Discrete) Visualizations:

- The **Late Months (3-6)** column has a majority of counts **(84%) at 0**, i.e. active accounts, with around11% of customers having at least 1 late count in the time, and a further 3% of customers have 2 counts
- The larger **Late Month columns (6-12, 12+)** have the large majority of counts **(95%) at 0**, i.e. active accounts, with around 3.4% - 3.8% of customers having at least 1 respective late count in that timeframe
- About **20% of the customers sampled are Late (1)** (blend of between 3 to 12 months or higher) and are viewed, for purposes of this analysis, as being **likely to default/non-renew**



Numeric (Continuous) Visualizations:

Distribution: **Normal (Mean Similar to Median):

- Customer Age
- Risk Score
- Number of Premiums Paid

Right/Positive Skewed (Mean Larger than Median):

- Percent Premium Paid Cash
- Avg. Monthly Premium
- Income/Premium

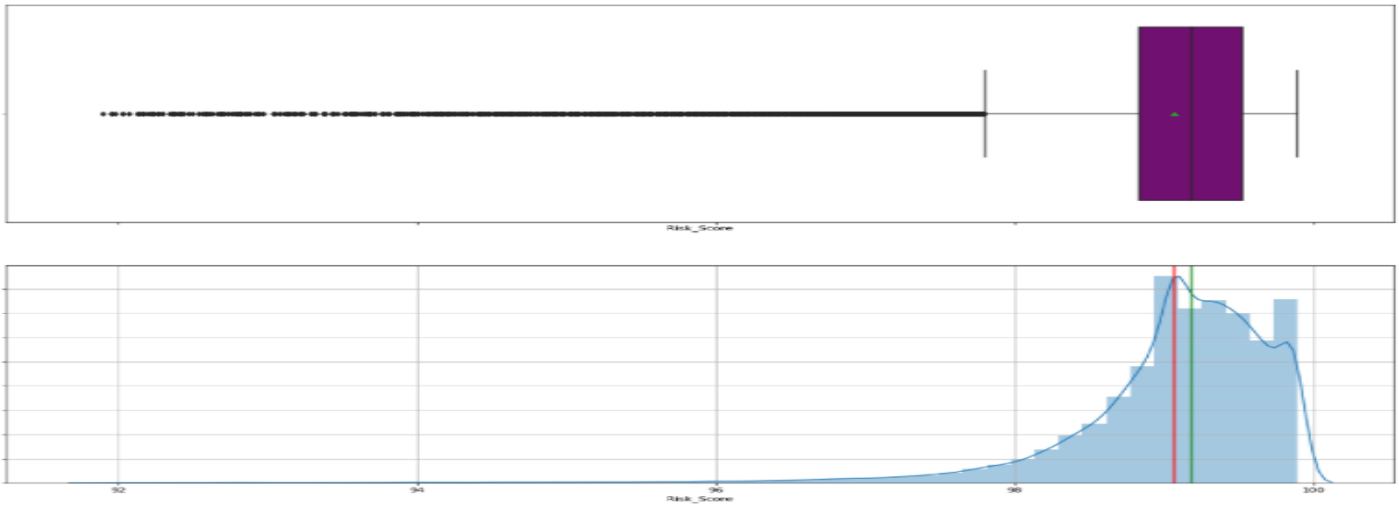
Outliers:

- **Customer Age:** Customers aged between 98 and 103 are higher than the majority, however very plausible nowadays and not assumed to be invalid data
- **Risk Score:** Scores on the lower/left side between 92% and around 97.6%, which appear valid within such a limited distribution range
- **Number of Premiums Paid:** Premiums, paid annually, on the higher/right side between 26 and 60 years, which appears valid when accounting for Customer Ages sampled, when assuming some customers paying into a given Life Insurance policy could be paying for decades (until death) to keep the policy active
- **(Annual) Premium:** Premiums ranging between around $28k and $60k each year do initially appear as outliers with data issues, however when analyzing and factoring in client net worth and subsequent Insurance Policies purchased, the annual premiums align proportionately

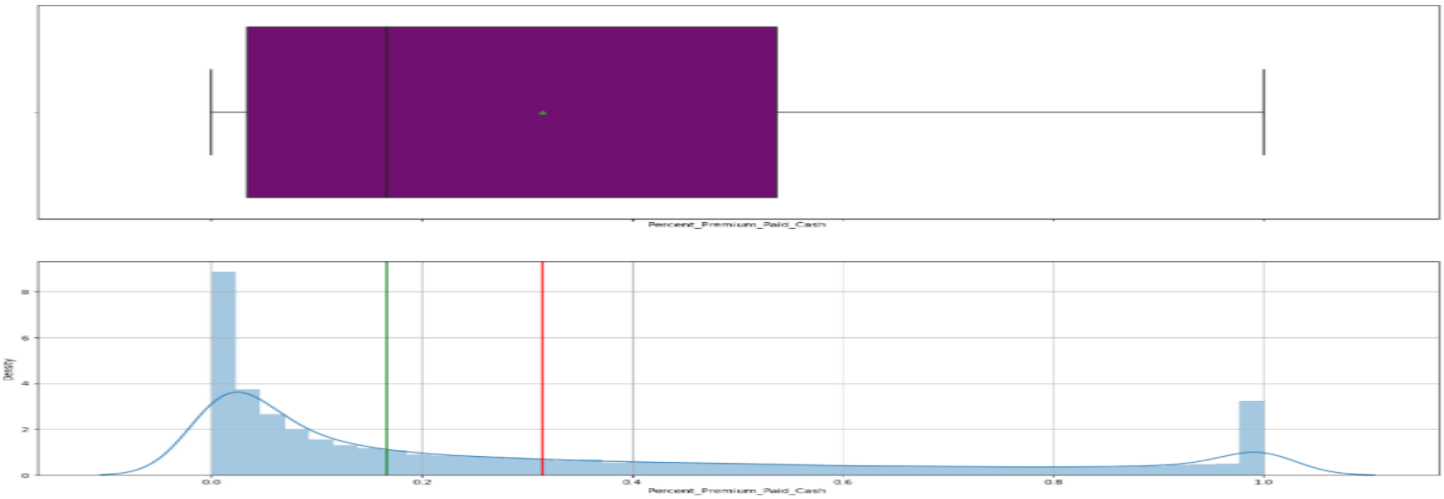
- **Avg. Monthly Premium:** Similar to Annual Premiums, those falling within the monthly equivalent of around $2.3k to $5k do initially appear as outliers with data issues, however as shown above are likely associated with higher net-worth clients willing to pay more for substantially higher coverage
- **Income/Premium:** Although the distribution has largely improved vs. the prior Income column due to the proportionate allocation against respective Premiums, there are numerous outliers in the range of around 25 to 1.5k times Income to Premium

Notable Distributions:

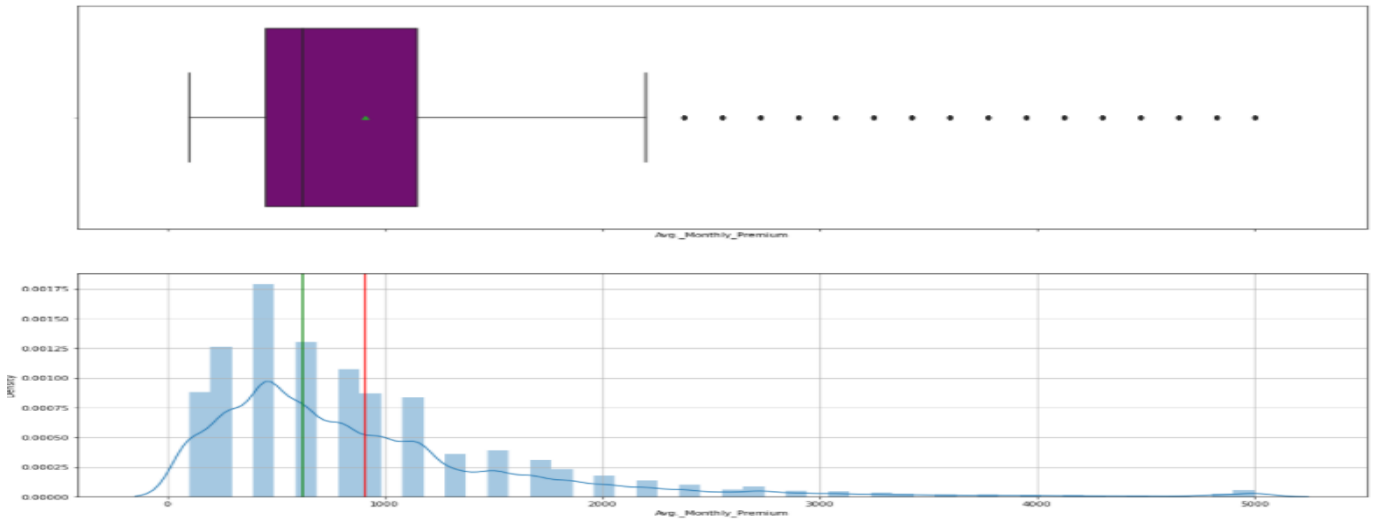
- Percent Premiums Paid Cash - **Right/Positive Skewed** (Mean Larger than Median)



- Risk Score - **Normal** (Mean Similar to Median):



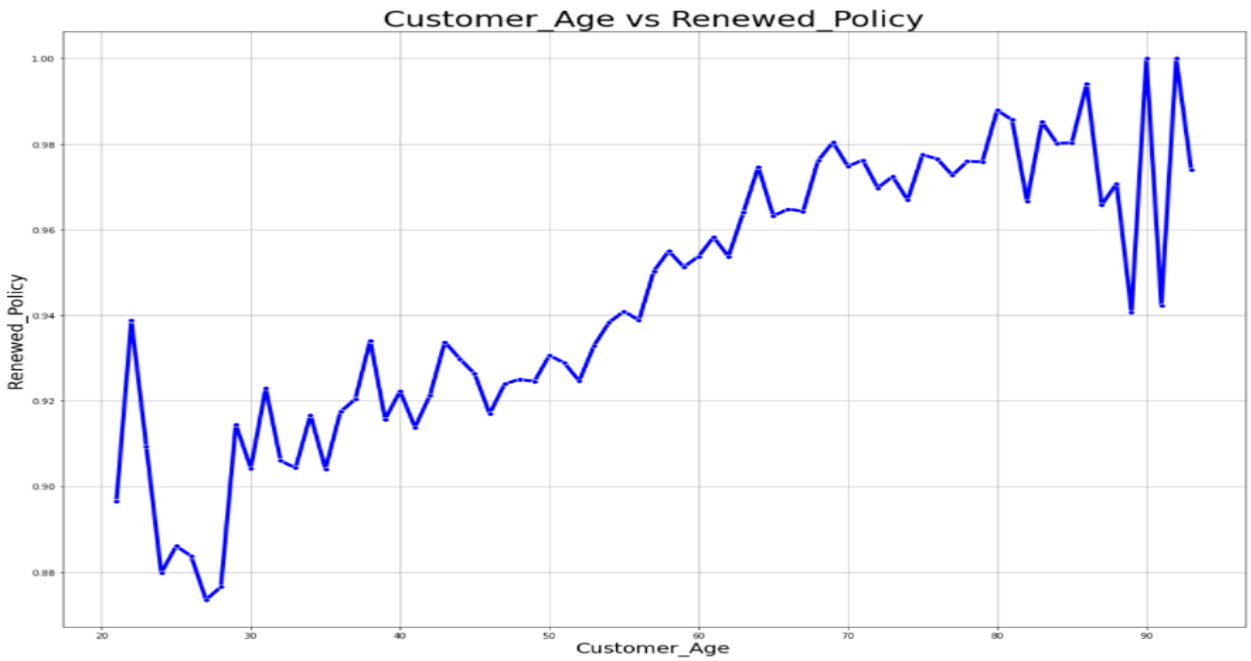
- Avg. Premium Paid - **Right/Positive Skewed** (Mean Larger than Median):



Bi-variate:

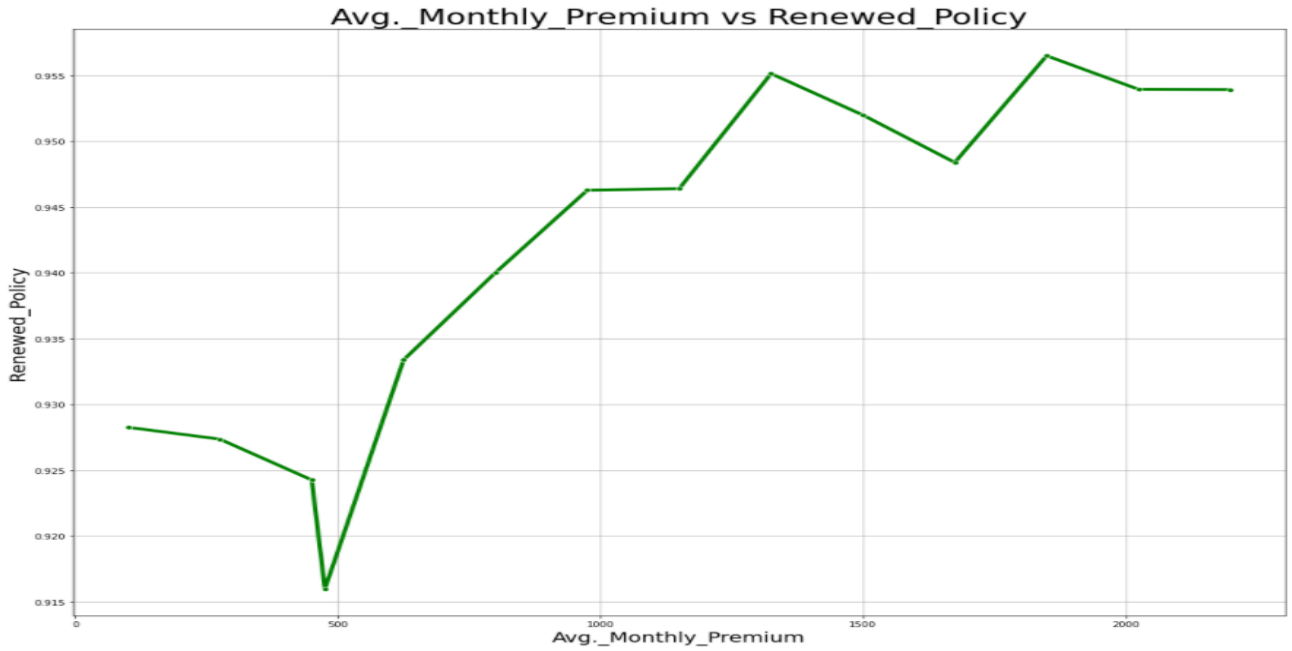
Renewed Policy by Customer Age:
The overall trend indicates that likelihood of Renewals substantially **increases, proportionately, as the age of customers increase** - i.e. negatively correlated to non-renewals (inverted).

- In general, **customers in the 30s to 40s range from 7% to 9% likelihood of Non-Renewal vs. customers nearing 70, with closer to 2% likelihood of Non-Renewal**
    - Customers in their late 80s to early 90s are less consistent in renewing their policies due to various factors, part of which could be the death of older customers no longer requiring continued insurance coverage and premium payments



Renewed Policy by Average Monthly Premium:

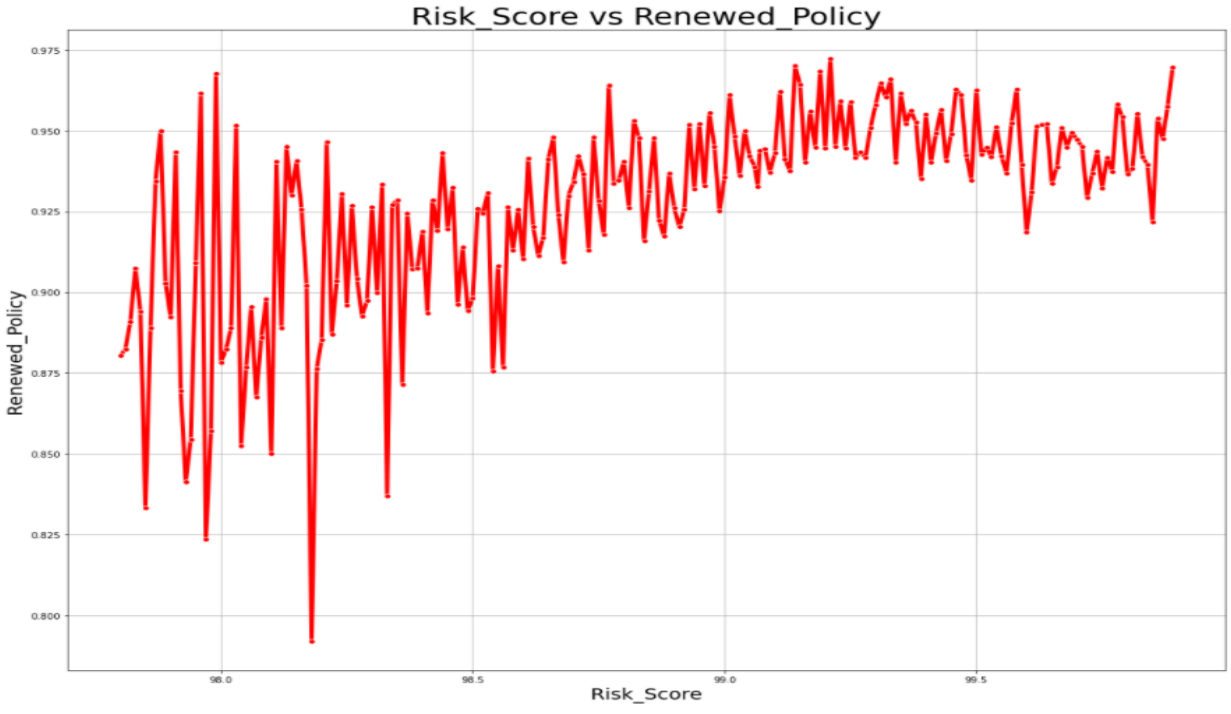
The overall trend indicates that, after a minimum payment of $500 a month, the **likelihood of Renewals improves as the Average Monthly Premium increases** - i.e. negatively correlated to non-renewals (inverted).



Renewed Policy by Customer Risk Score:

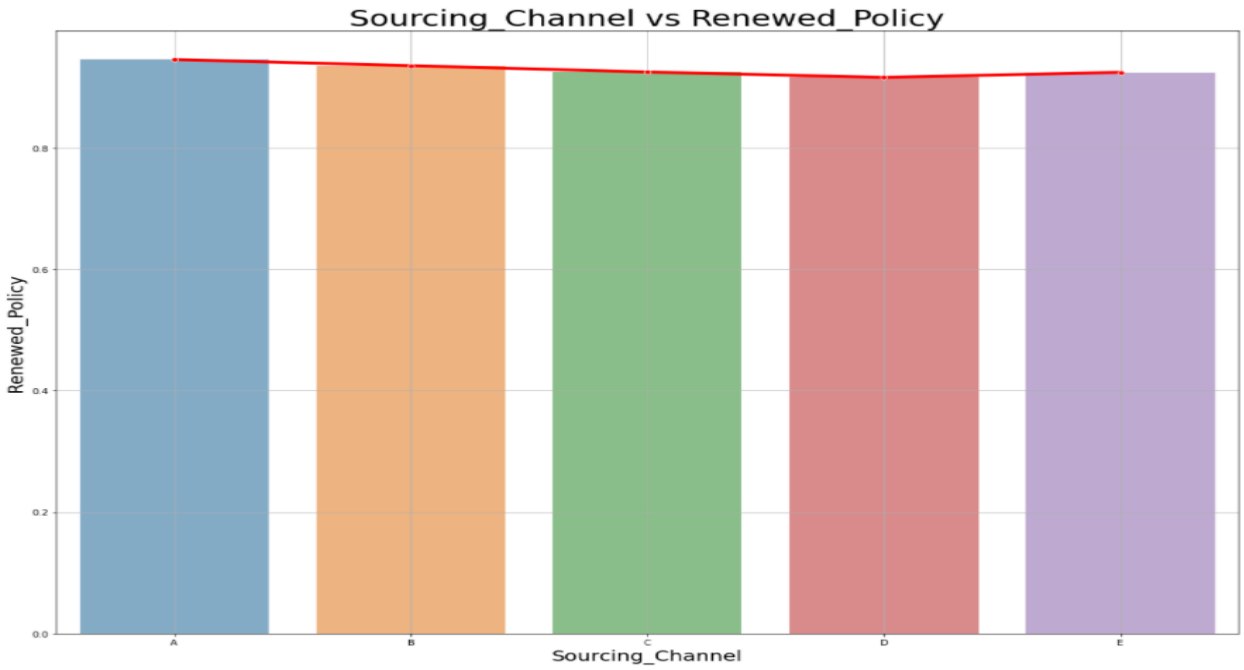
Less consistent in terms of correlation patterns, the general pattern indicates that **the likelihood of Renewal increases as customer Risk Scores improve** - i.e. negatively correlated to non-renewals (inverted). –

- A large reason for this is that higher risk customers are more likely to file a claim, hence the risk, and are **more incentivized to continue renewing their insurance policies**



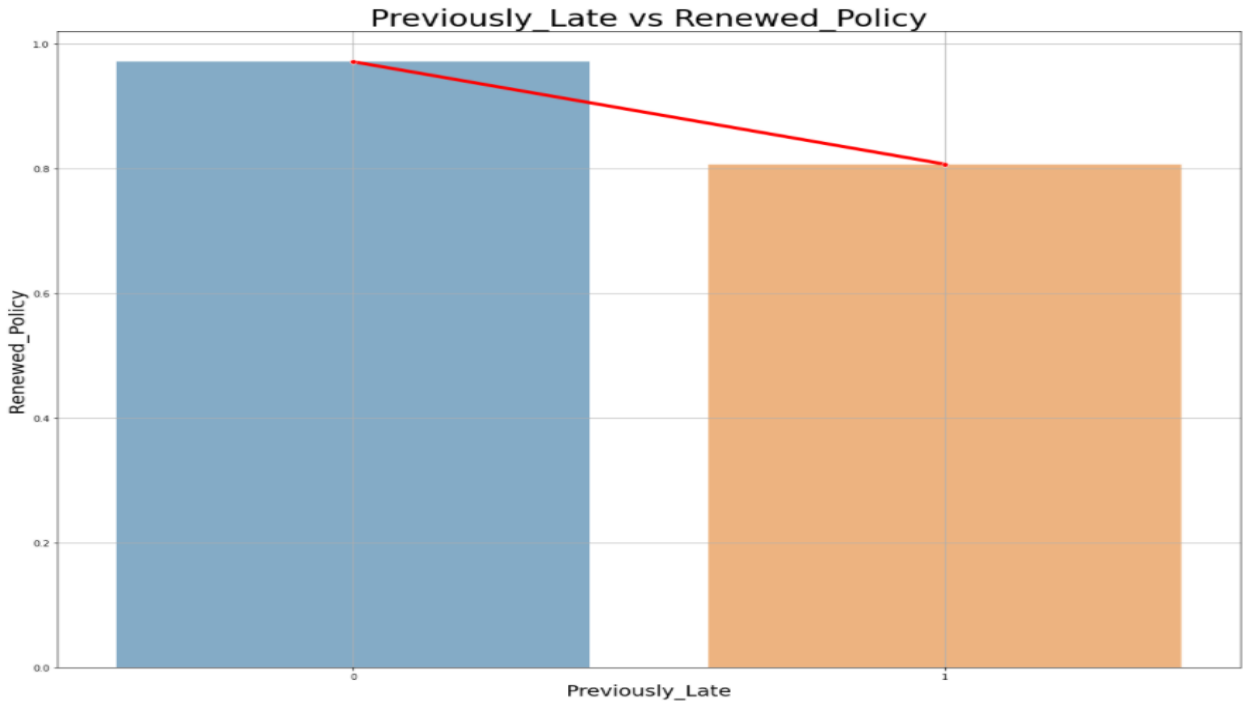
Renewed Policy by Sourcing Channel:

- We can see that **Channel D has the lowest renewals (highest split of customer non-renewals (8.4%))**, closely followed by Channels C and E with similar splits (7.5% and 7.6% non-renewals respectively)
- **Channel A has the highest renewals (lowest non-renewals (5.4%))**

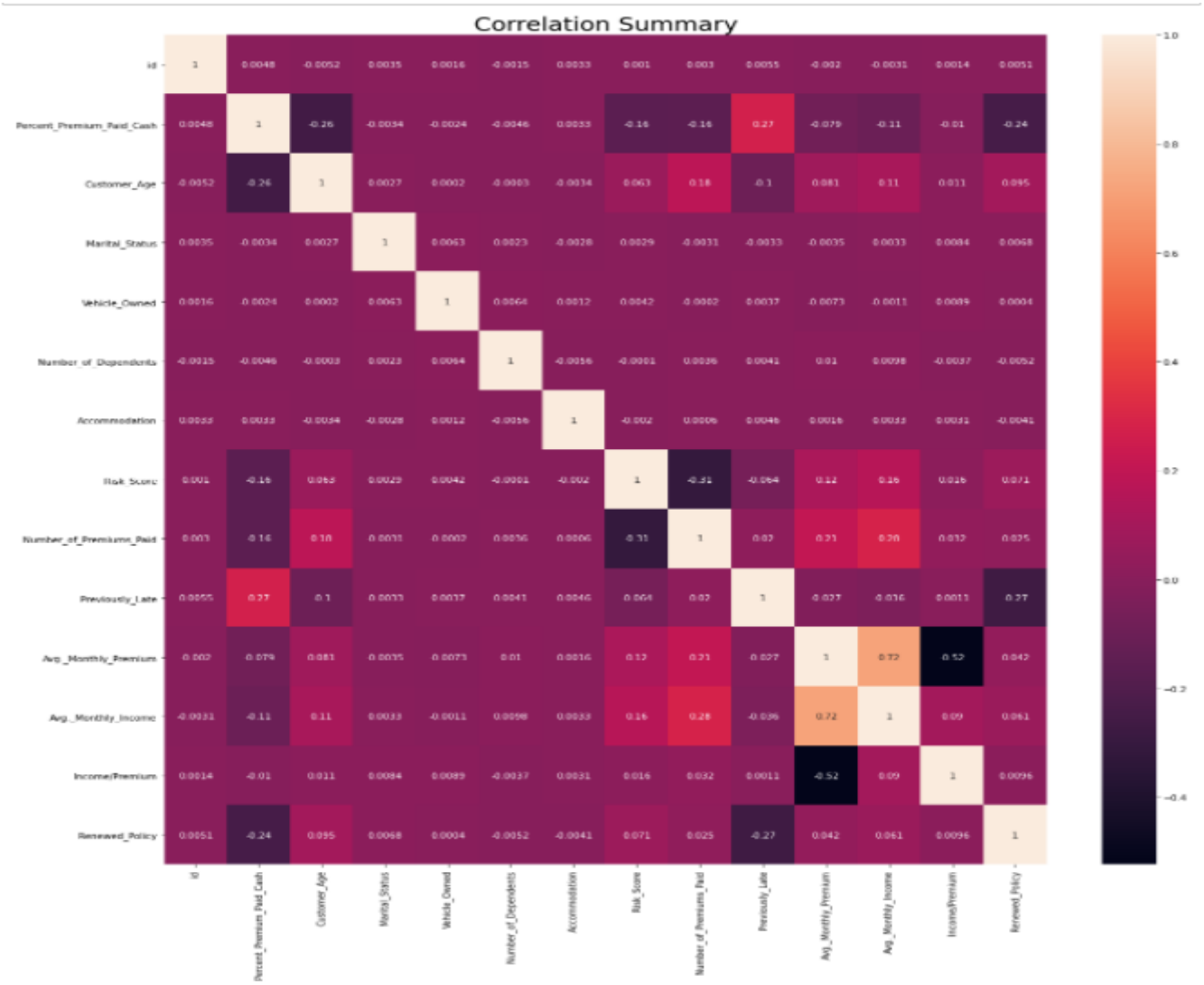


Renewed Policy by Previously Late Count:

- Customers who have been previously late over the course of their relationship with the business, account for upwards of 18% of the Non-Renewals (only 82% renew), **whereas those customers never late on premium only comprise around 3% of Non-Renewals**



Correlation:



- The strongest positive correlations, excluding Target variable, exist between:
  - **Positive Correlation**
    - Avg. Monthly Premium and Avg. Monthly Income (+72%)
    - Percent Premium Paid Cash and Previously Late (+28%)
    - Percent Premium Paid Cash and Avg. Monthly Income (+27%)
  - **Negative Correlation**
    - Avg. Monthly Premium and Income/Premium (-52%)
    - Number of Premiums Paid and Risk Score (-31%)
    - Percent Premium Paid Cash and Customer Age (-26%)
- Focusing on purely the Renewed Policy (Numerical Target) column:
  - **Positive Correlation**
    - Previously Late (-27%)
    - Percent Premium Paid Cash (-24%)

Additional Observations:

Risk Score:

- Indicative of likelihood of filing a claim, from a correlation standpoint
    - The entire sample distribution lies within a **small range (92%-99.9%) and is disproportionately assigning risk to all variables within the sample**

Percent Premiums Paid Cash:

- Unimodal, normal (Mean similar to Median), distribution
- Excluding Risk Score, no notable correlations to other variables sampled

Percent Premiums Paid Cash:

- **Bimodal distribution**, with a large portion of customers having paid less than half their policies in cash, however there is also a **notable segment of customers having paid upwards of 75% or more of their premiums in cash**
- Slightly increased likelihood of late count (blended 3-12+ months) on premiums closer to 75%+ paid in cash

Avg. Monthly Premium:

- **Mostly unimodal distribution**, accounting for outliers up to $5k a month for high net-worth clients, and is right skewed (Mean larger than Median)**
- **Mixed distribution in relation to Income/Premium ratio**, indicating that not all higher income customers have proportionately high premiums allocated to their account
    - These higher income customers may have other policies with other providers to split the risk allocation, etc.

Income/Premium:

- **Mostly unimodal distribution**, accounting for outliers up to $1.5k earned per Premium for high net-worth clients, and is right skewed (Mean larger than Median)**
- Somewhat negatively correlated to all Months Late counts (late counts increase as income decreases)
- **Largely unaffected by Marital Status, Vehicles Owned, Dependent Count, or Accommodation Type (excluding outliers)**

Data Pre-Processing

Missing Value Check/Treatment:

- **Null Values:** 0 null values found in sample dataset
- **NA Values:** 0 NA (NaN) values found in sample dataset
- **Duplicated Values:** 0 duplicate values found in sample dataset
- No treatment required for Missing or Duplicate values

Variable Transformation:

- **Sourcing Channel & Residence Area Type:** Datatypes converted from Object to Category
    - Reduces computational output and simplifies dataset functionality
- **Age in Days:** Renamed to Customer Age and converted to years by dividing by 365
- **Variables (Majority) Cleaned/Renamed, specifically:**

    - Percent_Premium_Paid_Cash
    - Customer_Age
    - Marital_Status
    - Vehicle_Owned
    - Number_of_Dependents
    - Accommodation
    - Risk_Score
    - Number_of_Premiums_Paid
    - Sourcing_Channel
    - Residence_Area_Type
    - Previously_Late
    - Avg._Monthly_Premium
    - Avg._Monthly_Income
    - Income/Premium
    - Renewed_Policy

Addition of New Variables:

- **Previously Late:** Combined sum of all three Month Late columns (3-6, 6-12, 12+), where anything over a **total of 0 is set to 1 (Yes) and all 0 counts are kept as 0 (No)**
- **Avg. Monthly Premium:** Premium column replaced by new feature, dividing each respective attribute by 12 for a monthly equivalent
- **Avg. Monthly Income:** Income column replaced by new feature, dividing each respective attribute by 12 for a monthly equivalent
- **Income/Premium:** Used as a relationship ratio between Income and Premium ratios, which were previously shown to have strong correlations, providing a much smaller attribute for analysis
- **Renewed Policy:** This feature replaced the renewal column, with was originally numeric (0/1), to an equivalent categorical column (No/Yes)
    - This made for cleaner EDA to follow and also reduced the required steps to later convert and invert the attributes for targeting the Minority Class as 1 (i.e. Default/Non-Renewal)

Removing Unwanted Variables:

- **Count (3-6 Months Late):** Confusing to interpret when combined with other monthly counts – replaced by single collected column, Previously Late
- **Count (6-12 Months Late):** Confusing to interpret when combined with other monthly counts – replaced by single collected column, Previously Late
- **Count (More than 12 Months Late):** Confusing to interpret when combined with other monthly counts – replaced by single collected column, Previously Late
- **Premium:** Annual totals replaced by Monthly Premium equivalent, Avg. Monthly Premium
- **Income:** Annual totals replaced by Monthly Premium equivalent, Avg. Monthly Income
- **Renewal:** Replaced by categorical (No/Yes) equivalent column, Renewed Policy

Outlier Identification:

- Initial Box Plot graphs created for numerical (continuous) values in dataset to identify outliers for each feature
- Outliers found for the following features:
    - **Customer Age:** Customers aged between 98 and 103 are higher than the majority, however very plausible nowadays and not assumed to be invalid data
    - **Risk Score:** Scores on the lower/left side between 92% and around 97.6%, which appear valid within such a limited distribution range
    - **Number of Premiums Paid:** Premiums, paid annually, on the higher/right side between 26 and 60 years, which appears valid when accounting for Customer Ages sampled, when assuming some customers paying into a given Life Insurance policy could be paying for decades (until death) to keep the policy active
    - **Avg. Monthly Premium:** Similar to Annual Premiums previously reviewed, those falling within the monthly equivalent of around $2.3k to $5k do initially appear as outliers with data issues, however as shown above are likely associated with higher net-worth clients willing to pay more for substantially higher coverage

- **Avg. Monthly Income:** Just as Avg. Monthly Premiums include numerous outliers to the right of the distribution, the Average Monthly Income includes the same, making for very hard to comprehend analysis of data points
- **Income/Premium:** Although the distribution has largely improved vs. the prior Income column due to the proportionate allocation against respective Premiums, there are numerous outliers in the range of around 25 to 1.5k times Income to Premium

Outlier Treatment:

- Function created to calculate the **IQR (area between 3rd and 1st quantiles)**
- A multiple of 1.5x the IQR is applied to whichever end the outlier/s exist, namely:
  - Subtracted from the 1st quantile for features with lower end outliers (e.g. Risk Score)
    - **An additional limit of 0 applied for features where lower bound IQR less than 0**
  - Added to the 3rd quantile for features with upper end outliers
- With the resulting Outlier treatment applied, the new Minimum and Maximum ranges are as follows:
  - **Customer Age:** 21 - 93
  - **Risk Score:** 97.8% - 99.9%
  - **Number of Premiums Paid:** 2 - 24
  - **Avg. Monthly Premium:** $100 - $2,200
  - **Avg. Monthly Income:** $2,003 - $39,017
  - **Income/Premium:** 4.2x - 42x

It is also worth noting that, although the maximum values for Average Monthly Income and Premiums may still appear high initially, all are adequately correlated to higher income/premium customers and are therefore accurately presented.

Additional analysis of the sample set, coupled with the substantially upper bounds provided by insurance products, indicates that these are indeed **higher priced premiums for higher net-worth/earning individuals to cover their lifestyle needs accordingly.**

Encoding & Data Splits (Train/Validation/Test):

Label Encoding Features:

- Sourcing Channel converted from Categorical feature to numeric through manually label encoding (ordinal) channels A-E to 0-4 respectively

X/y & Train/Validation/Test Splits:

- **Split 1 – X (Independent Data) and y (Dependent Target):**
  - X set to all features excluding the target, Renewed Policy
    - Additionally during this step, Renewed Policy categorical column dropped as no longer needed (redundant)
  - y set to the target feature, Renewed Policy
- **Split 2 – X Temp/Test & y Temp/Test:**
  - **To avoid Data Leakage**, where Training data testing/updates made could lead to biased changes in Test data), an additional step of using Validation data is used, initially creating a Temp dataset
    - The Stratify step is used so as to keep the same X/y proportions throughout the splits

- **Split 3 – X Train/Validation & y Train/Validation:**
    - Following up on the previous step, the **Temp dataset is now converted to Validation**
        - **X Train = 75% of X Temp** (which is taken as 80% of X initially)
        - **X Val = 25% of X Temp** (which is taken as 20% of X initially)
        - The Stratify step is again used during this step, so as to keep the same X Temp/y Temp proportions throughout the splits
- All updates and testing should only occur between the Train and Validation datasets, **with final model testing applied at the very end to the Test dataset.**

One-Hot Encoding Features:

- Using Pandas Get Dummies feature to one-hot encode Residence Area Type to numeric
    - This creates a new column for each unique attribute within the feature name with a corresponding 1 (+1 to feature count) or 0 (+0 to feature count) reflected on each row for each respective attribute
    - An additional option was selected to remove the first additional column added and, since there were originally only two columns added, the net change was 0
        - The logic behind removing the first column is that the model intuitively knows that if based on the other feature results (e.g.0 or 1 for 2 columns), we can infer the other result accordingly

Scoring the Model Results:

In addition to the overall model accuracy score (% correct vs. incorrect), it is also important to consider the following score metrics for the Classification model:

Without reversing the target classes (majority = Renew/1), the model will predict based on likelihood of renewal vs. non-renewal, using the following metrics:

- **Precision:** How many of the customers predicted to renew their policies actually did?
    - True Positives / (True Positive + False Positives)
- **Recall (Sensitivity):** Of all the customers that renew their policies, how many did the model predict would?
    - True Positive / (True Positives + False Negatives)
- **F1-Score (Harmonic Mean of Precision & Recall/Sensitivity):** What is the Harmonic Mean split between the Precision and Recall results?
    - 2 * (Recall * Precision) / (Recall + Precision)
- **Specificity:** Of all the customers that did not renew their premiums, how many did the model predict wouldn't?
    - True Negative / True Negatives + False Positives

Reducing Loss and Accounting for Imbalanced Data: Specificity

- Due to the fact that the company isn't as much focused on predicting for positive cases (Renewals) but rather accounting for all possible negative cases (Non-Renewals), **Specificity will be the score to optimize for from a predictive modelling standpoint**
    - However, Recall and Precision are also important for further inspection and analysis (particularly during Confusion Matrix summarization)
        - **Recall is also important in identifying and limiting customers predicted not to renew that actually end up renewing**, as this can cost the company wasted time, resources, and opportunity costs, in addressing non-risk clients over those truly at risk of default

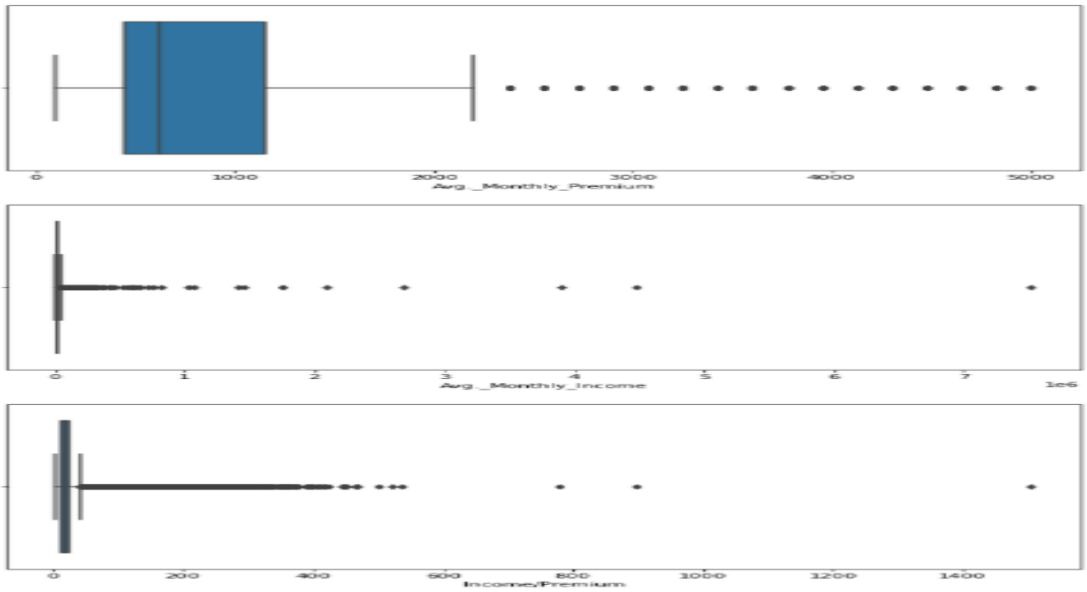
- It is good to ensure that the model is still performing high on Precision and correctly predicting the maximum amount of customers expected to renew their policies, likely to be well above 90% due to data imbalance (94% of data included renewals)

EDA (Post Pre-Processing

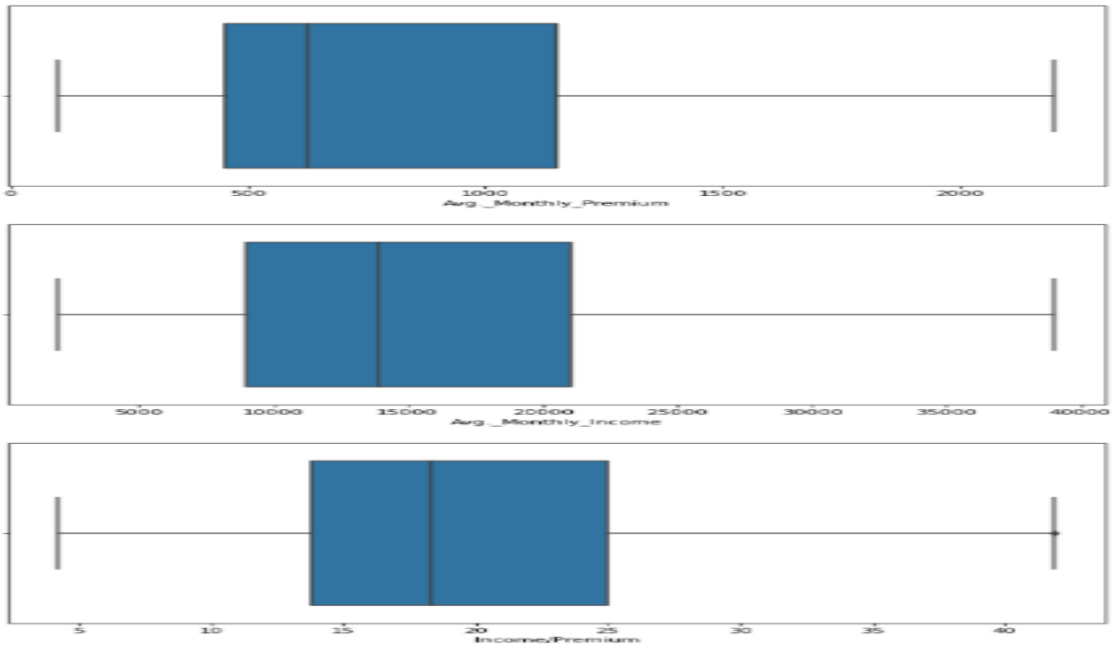
Relationship Amongst Variables:

- Following successful outlier treatment during Pre-Processing state, all numerical (continuous) variables previously skewed by outliers are now interpretable, most notably related to Income and Premium:

Before Preprocessing/Outlier Treatment:



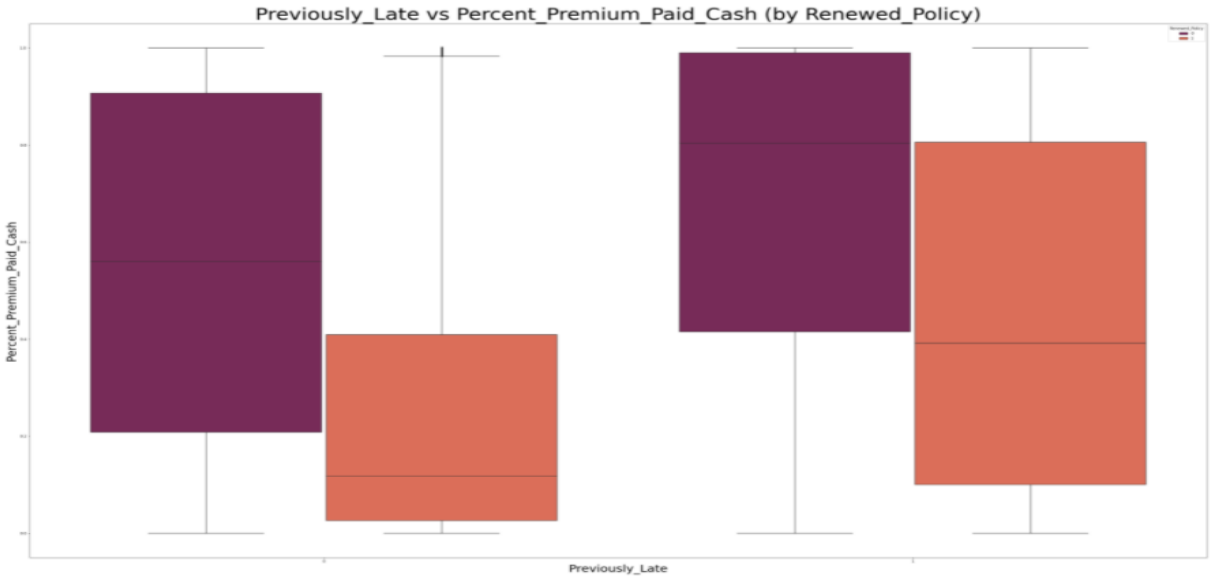
After Preprocessing/Outlier Treatment:



Important Variables (Interaction Analysis):

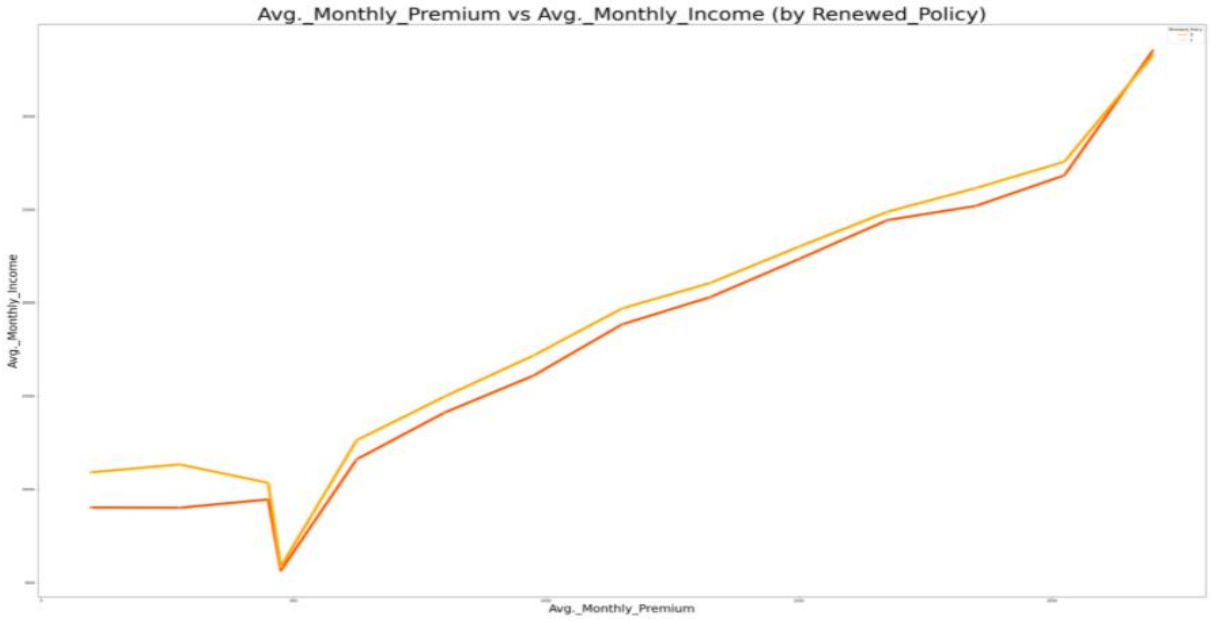
Percent Premium Paid Cash & Previously Late (Renewed Policy Split):

- In general, customers that have **paid higher portions of their policy with cash as opposed to credit, are far more likely** to have been, or will be, late on their accounts, as well as potentially defaulting (non-renewing) their policies
- Similarly, customers who have never previously been late on their account, and have paid less of their current policy with cash, are the most secure and least likely to default or non-renew in the future



Avg. Monthly Income & Average Monthly Premium (Renewed Policy Split):

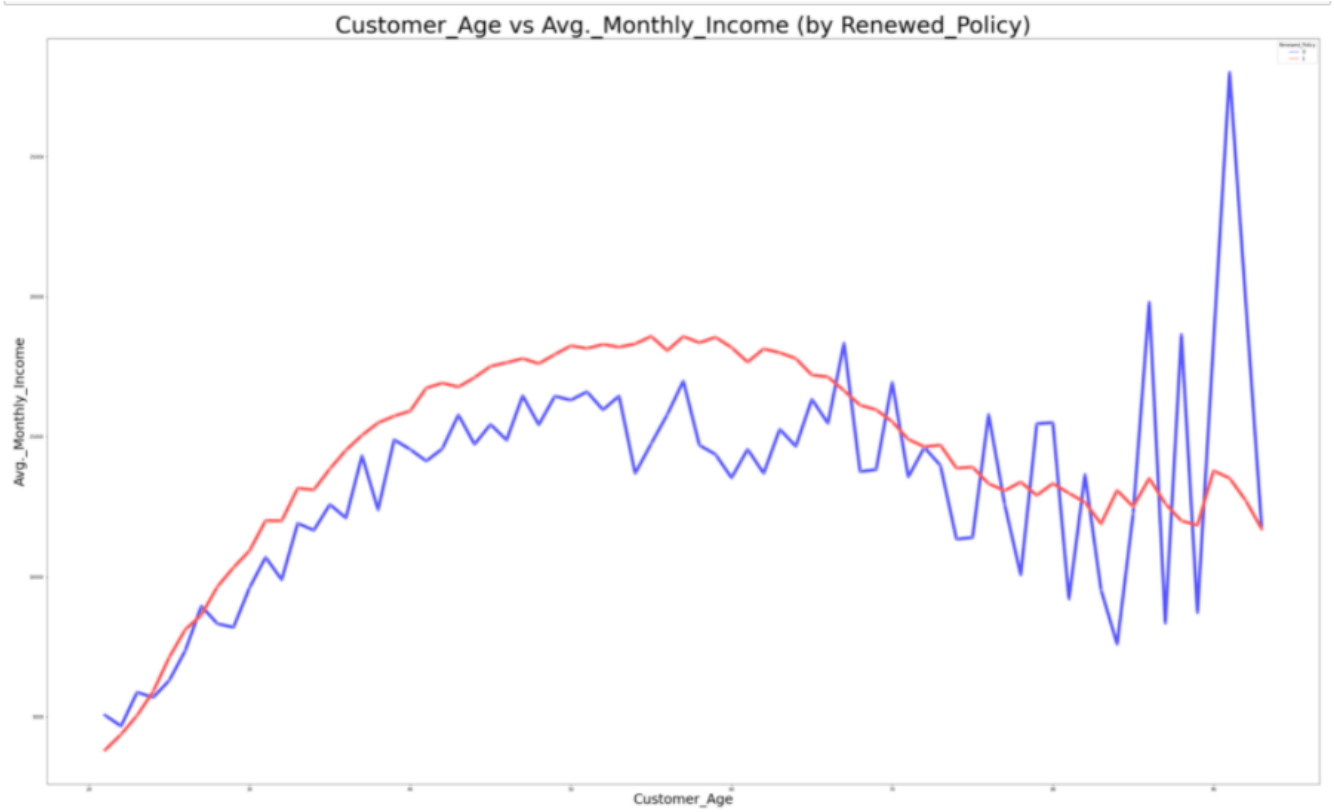
- As was expected, the **average monthly premiums paid correlate strongly with respective average monthly incomes, increasing proportionately**
- The gap between renewals/non-renewals appears larger (less proportionate) for customers earning under $10k a month and paying under $500 a month for their policies



Insight Visualizations:

Customer Age and Avg. Monthly Income (Renewed Policy Split):

- **Average monthly income increases with age, up to a point from around 50 to 60 (where some customers opt to retire or work less)**, before dropping as is somewhat expected when individuals transition out of full-time employment income sources into retirement funding
- There is a large spike in average monthly income, as well as Renewed Policy, for some customers in their mid-80s through mid-90s
    - This could be **indicative of customers cashing out various (final) retirement products and making the decision to opt out of future policy payments** that are deemed unnecessary for covering remaining life expectancy



Alternative Analytical Approach

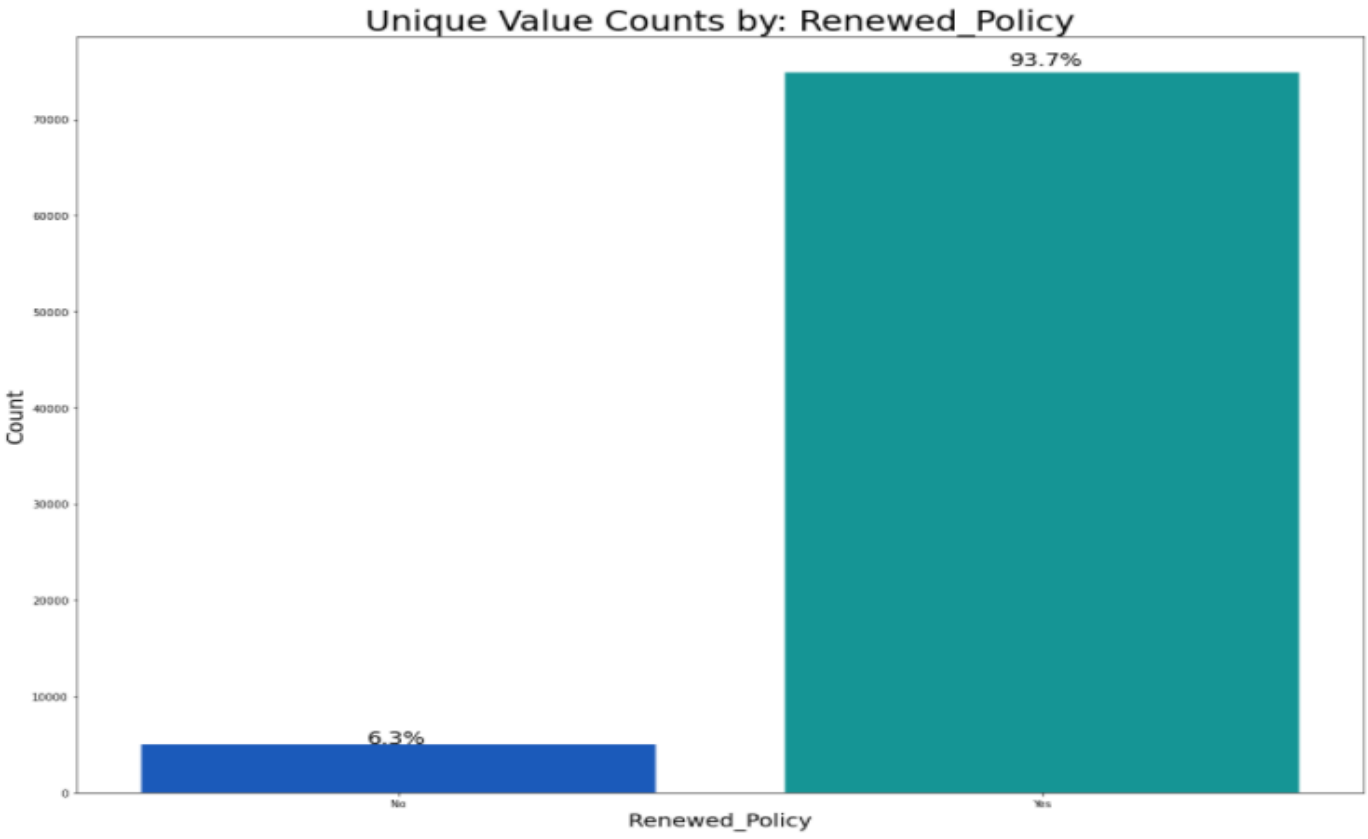
Flipping Target Variable Counts Early – Set Target (Minority) to 1:

Instead of only flipping the target counts in the Renewal column during final EDA ahead of the Train/Validation/Test split, I also investigated switching the numerical values of 0 and 1 for respective Minority and Majority classes from the onset of the analysis, within the Renewal column.

I found, however that there was initially some benefit in seeing which customers had or had not renewed their premiums and the largely imbalanced split that existed within the sample, without a specific target (minority) utilized in driving the analysis.

What I did initially do, however, was to **rename the column to Renewed Policy convert the numerical values to categorical as No (0) or Yes (1)**.

This sufficiently showed the imbalance in data, which would substantially affect prediction model performance and **required approach to combat it (Up-Sampling and Scaling data, etc.**



Imbalanced Classes: Majority (0) vs. Minority (1):

I tested inverting the classes so that the target was set to 1 for Default (Non-Renewal) instead of 0 as Non-Renewal and ran preliminary prediction models on those to test initial results (Accuracy and Recall) primarily. I found that the results were very low simply due to the fact that on average 6% of the time the model would be right, based on the sample split of 6% Non-Renewal to 94% Renewal, for current customers sampled.

I quickly found that this substantially skewed dataset would not need to be inverted and instead could be left as is, with the Confusion Matrix metrics (Specificity for targeting Non-Renewals for example) utilized for segregating scoring based on business goals – targeting at-risk customers not expected to renew.

All Bivariate and Interaction Analysis (Multivariate to Target) were updated and reexamined against the **updated target column, Renewed Policy**.

Modeling Process

Train/Validation/Test Splits

The Customer ID column, originally serving as an index column, was not dropped, as it offers a unique identify for each customer sampled. Instead, it was **set to the index of the X dataframe (not counted as a column),** and will be easy converted back to a column (reset) when needed for identifying specific customers in the predicted results dataset.

To further protect the Test dataset from data leakage, an addition Validation dataset was created for initial model testing, with final approved models being fitted to Test data only at the very end of the analysis.

Final data splits resulted in the Training dataset comprising of 60% of the X data, with the Validation and Test datasets each comprising of 20% respectively.

Feature Encoding

- Sourcing Channel was manually encoded to a respective numerical (ordinal) equivalent – A:E became 0:4
- Residence Area Type was One-Hot Encoded to binary, with the first column dropped as to avoid redundancy in the dataset (if not 0, then 1, etc.)

Building the model: Evaluating Model Results

As previously mentioned, the company wants to identify all customers at risk of non-renewal, therefore needing to **accurately predict all Negative values and limiting False Positives** (predicted to Renew only to Default/Non-Renew).

Specificity is a crucial metric for predicting True Negatives **in time for early action so as to improve the renewal rate of at risk customers.** Due to the large class imbalance (6% Non-Renewals / 94% Renewals), achieving a very high Specificity score may not be possible**, and additional attention to** Precision **(correctly predicting Renewals) and** Recall **(limiting False Non-Renewal predictions which could be costly) need to also be analyzed as a cohesive summary.**

Building the Pipeline - (Upsampling/Scaling/Classification)

Need for Over/UnderSampling Data (Random Over/Under Sampler & SMOTE)

Due to the extreme imbalance between the classes of data (0: 6%, 1: 94%), it is necessary to add sampling techniques to upsample the minority class to equal the majority class within all datasets (Train/Validation/Test).

- Random Over Sampler tends to overfit the dataset for Training data, and if found to be substantially overfit should be replaced by Synthetic Minority Oversamping Technique (SMOTE), which uses the KNN algorithm to create synthetic training examples that increase variety within the dataset
- Due to the imbalanced dataset, the minority class will not be equally balanced to 50/50, which is highly unrealistic, but rather increased to a 20/80 split
    - This ratio can be adjusted after further threshold (ROC AUC Curve) testing

Need for Scaling

Although the majority of outliers were addressed and capped back to their respective Lower/Upper bounds plus 1.5x the IQR, **it is not sufficient in dealing with variances between unit metrics within the dataset (e.g. Risk Score = %, Income/Premium = $**, etc.). As a result, we need to further scale the data utilizing various methodologies including:

- **Standard Scaler:** Features scaled down to their respective Means and number of Standard Deviations from the Mean
- **MinMax Scaler:** Features are converted to a range of 0 to 1, based on the respective attribute's ratio of min/max as a percentage)

Use a Pipeline for All Steps (Simultaneously)

By simultaneously running the OverSampling steps with the scaling process, then running various Cross Validation and scoring procedures, the variances between each CV fold's results will be less and the model more generalized and consistent. Pipelines also make for cleaner and more reliable processing of multiple model steps, applied across various models.

Models & Data Pipelines:

The training dataset was run through a pipeline and upsampled using Synthetic Minority Oversampling Technique (SMOTE), utilizing 10 k-nearest neighbors and a sampling **strategy of 0.5 (Minority class (0) increased to be half of Majority class (1)**, or 33% / 67% split.

Pipelines built out for each model included the Standard Scaler (standard deviations from mean) for Logistic Regression, with other model pipelines only comprised of their respective classification models, initially. The models selected are:

- Logistic Regression
- Decision Tree Classifier
- Bagging Classifier
- Random Forest
- Adaptive Boost Classifier
- Gradient Boost Classifier
- XG Boost Classifier

Interpretation and Validation:

Each model pipeline was fitted to the upsampled Training data, with **initial ROC AUC scores calculated and Curves plotted, with optimal Thresholds identified**, for both the Training and Validation datasets. Variances between the scores for both sets were compared to ensure relative generalization within each pipeline (no obvious overfitting), with the **top 3 most generalized models selected for further detailed analysis**, including Confusion Matrix summaries, focused around Specificity in addition to general strong metric performance, and targeted hyperparameter tuning for additional performance improvements

Model Comparisons:

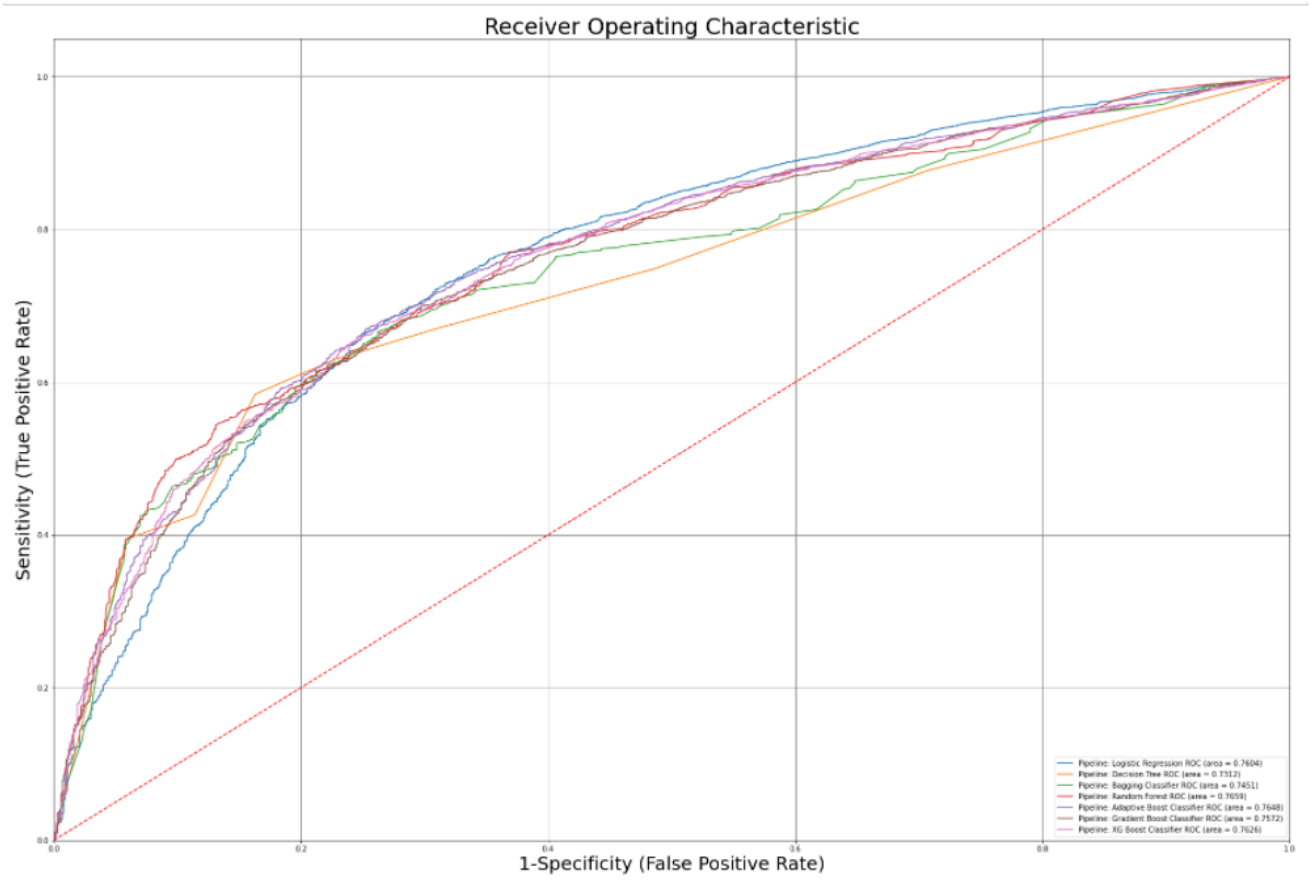
Roc Curves:

The **Receiver Operator Curve** and the calculated **Area Under the Curve (between 0 and 1)** can be very helpful indicators in identifying which models are best performing on various sample sets (Training and Validation in this case).

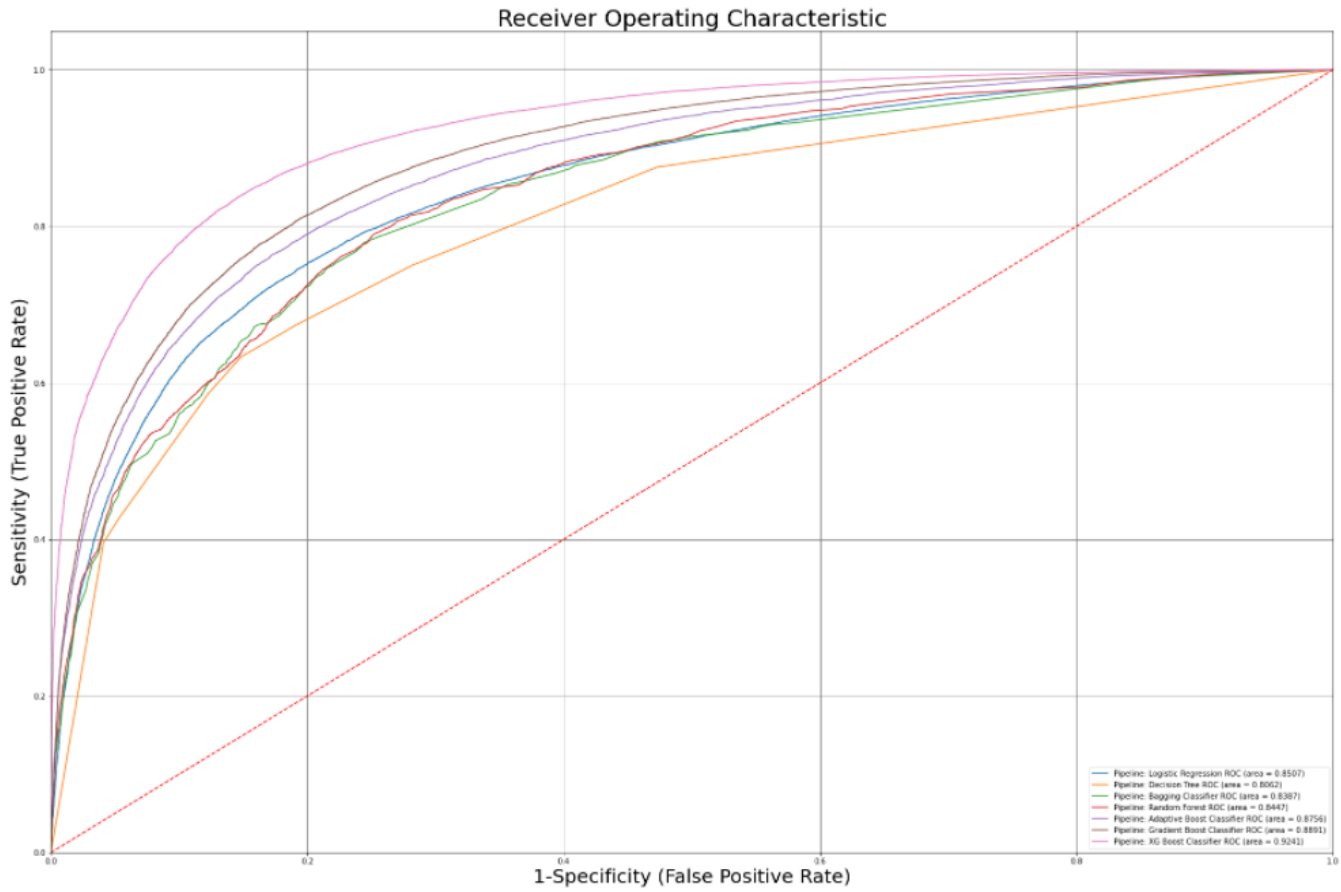
On the Y-Axis lies the True Positive rate (between 0 and 1) and on the X-Axis lies the False Positive Rate (also between 0 and 1). The closer the curve of the model lies to a score of 1 on Y-Axis (and a score of 0 on the X-Axis by relationship), the better a model is expected to perform.

ROC Curves were plotted for the 7 pipeline models, for both the Training and Validation data with the goal of selecting the top 3 most generalized models – scores with lowest variance between Training and Validation results.

Training Data



Validation Data



All models scored relatively well in general (approximately 82%) on the AUC scores for Validation Data. The specific Training and Validation scores will be calculated below and the top 3 generalized models (similar scores for both datasets) will be selected.
The models that are the **most generalized (small variances in probabilities for majority class)** for Training and Validation datasets are:

ROC AUC Scores - Training vs. Validation Date:

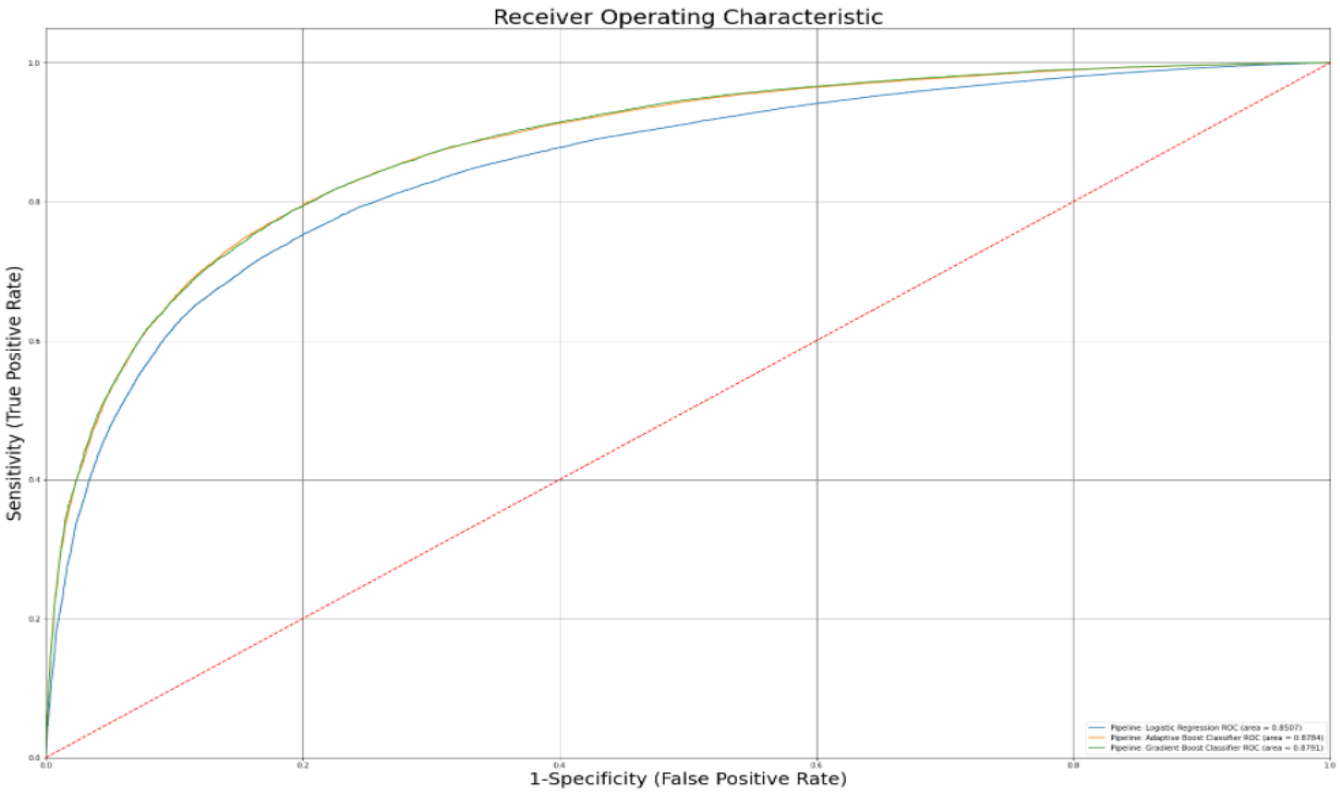
- Pipeline: Logistic Regression (85% vs. 82.4%)
- Pipeline: Adaptive Boost Classifier (87.7% vs. 82.1%)
- Pipeline: Gradient Boost Classifier (88.7% vs. 82.5%)

The remaining models may have scored higher for Training data probability predictions, but due to the disconnect with Validation data results, can be removed from further analysis going forward.

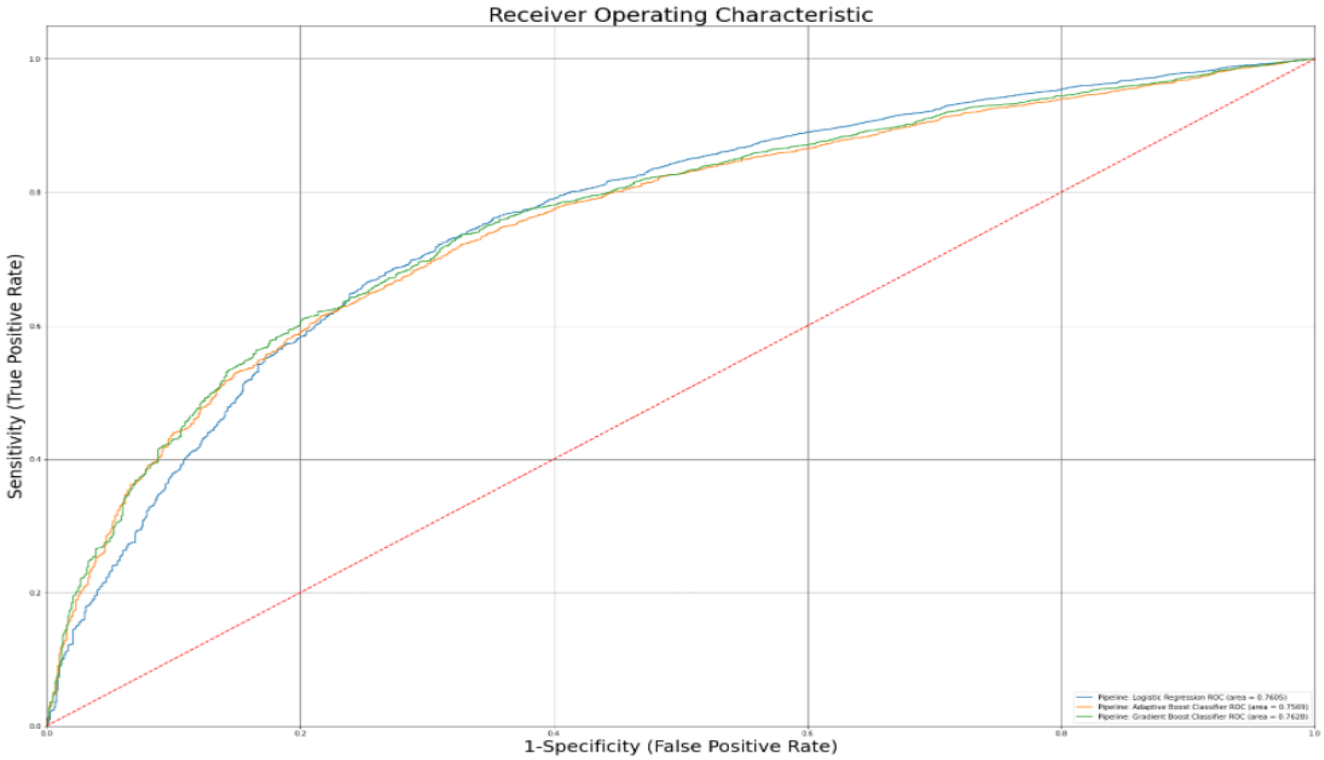
Final 3 Models – Post Tuning

Roc Curves:

Training Data



Validation Data



After tuning the top 3 models with the Best Parameters from the Grid Search CV review, the ROC AUC scores are:

ROC AUC Scores - Training vs. Validation Date:

- Pipeline: Logistic Regression (85% vs. 82.4%)
- Pipeline: Adaptive Boost Classifier (87.9% vs. 82.1%)
- Pipeline: Gradient Boost Classifier (87.6% vs. 82.4%)

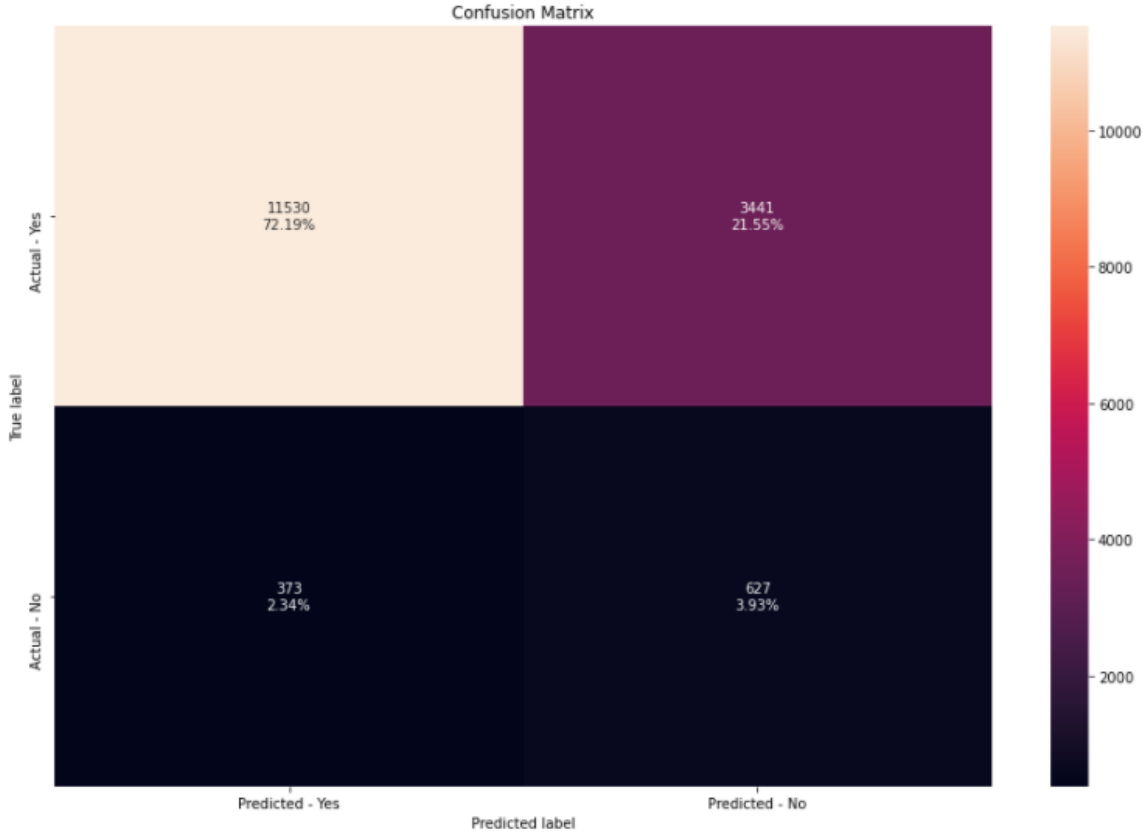
Although the scores are less generalized than before hyperparameter tuning, the primary goal is to improve Specificity scores which further Confusion Matrix reporting should confirm.

Confusion Matrix – Validation Data

Pipeline: Logistic Regression

```
Accuracy: 0.7612
Precision: 0.9687
Recall: 0.7702
F1 Score: 0.8581
Specificity: 0.627
.................................
Roc_Auc_Score: 0.7605

TP: 11530, FN: 3441, TN: 627, FP: 373
```



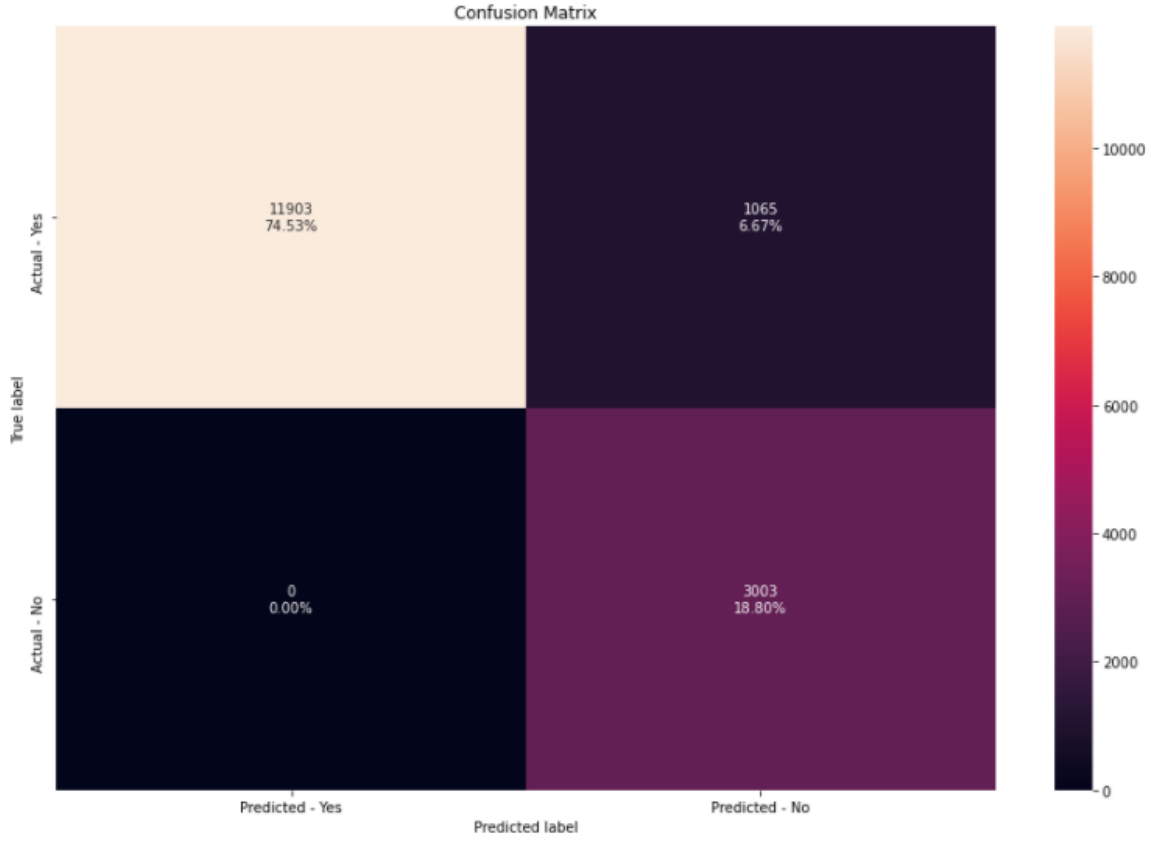
- The model scored very high for **Precision (True Positives / (True Positives + False Positives)** and relatively well for the target Metric, **Specificity (True Negatives / (True Negatives + False Positives)**
- Being that the primary objective for the business is to minimize False Positives (predicted to Renew but actually Defaulting/Non-Renewing), **a False Positive score of 2.3% is strong**

- Although the model scored lower in Recall/Sensitivity (True Positives / (True Positives + False Negatives)), in this business scenario that simply equates to customers predicted to Not Renew but actually Renewing
    - **The overall False Negative count (21.6%) can be improved upon through Adjusted Probability Thresholds**
    - As long as this miss doesn't result in large amounts of resources wasted unnecessarily on targeting customers not truly at risk, it can be seen as a slight benefit to the business, with the key focus being targeting true at risk customers (likely only 6% of the population based on sample splits)

Pipeline: Logistic Regression – Threshold of 0.4

```
Accuracy: 0.9333
Precision: 1.0
Recall: 0.9179
F1 Score: 0.9572
Specificity: 1.0
.................................
Roc_Auc_Score: 1.0

TP: 11903, FN: 1065, TN: 3003, FP: 0
```

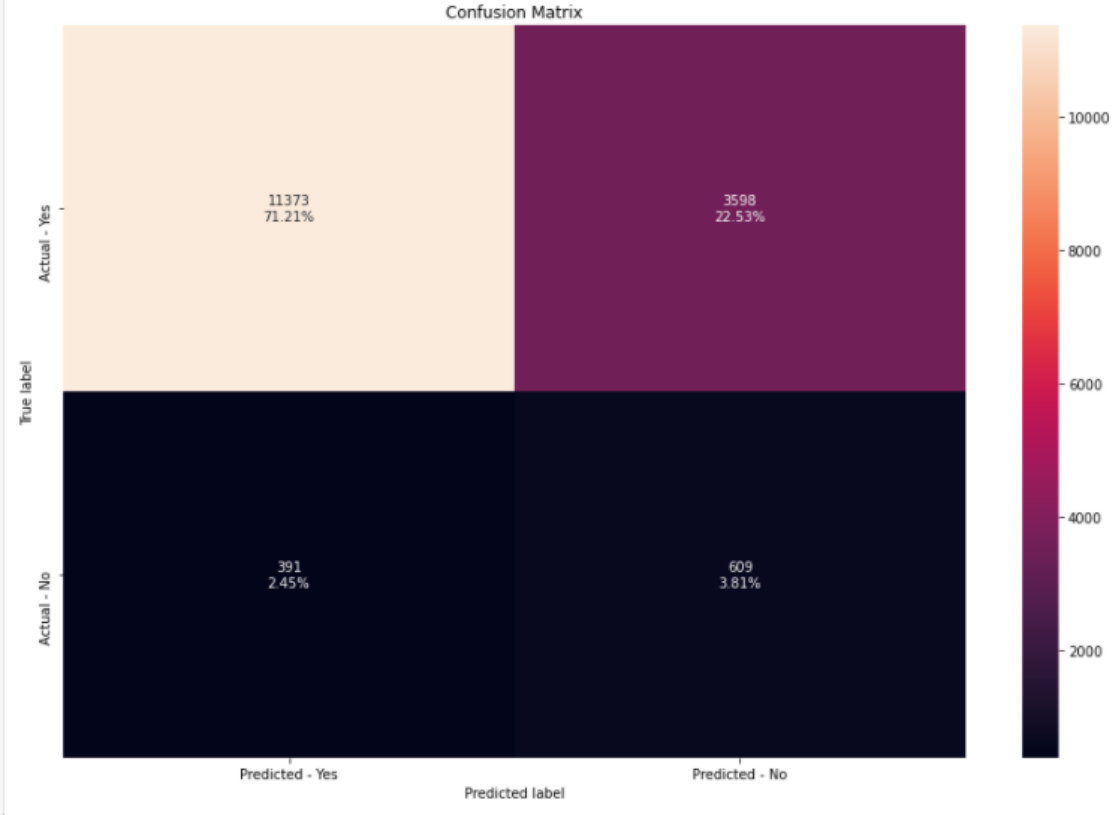


- **Lowering the threshold to 0.4** (vs. default of 0.5), after which point a customer is predicted to Renew, **lowers the False Negative Count from 21.6% to 6.7% and lowers the False Positive Count to 0%**
    - This results in perfect scores for Precision and Specificity, and substantially improves Recall (91.8%)

Pipeline: Adaptive Boost Classifier

```
Accuracy: 0.7502
Precision: 0.9668
Recall: 0.7597
F1 Score: 0.8508
Specificity: 0.609
.................................
Roc_Auc_Score: 0.7555

TP: 11373, FN: 3598, TN: 609, FP: 391
```

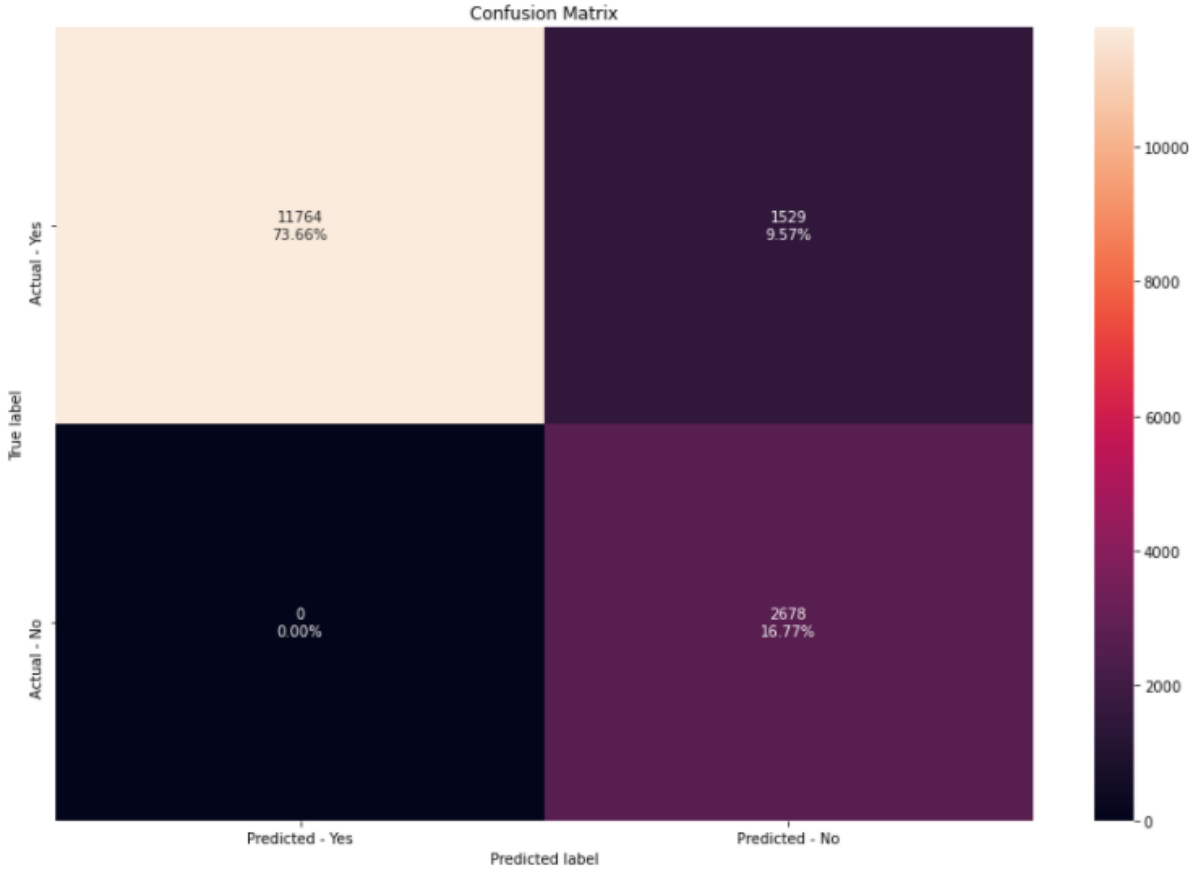


- The model scored very high for **Precision (True Positives / (True Positives + False Positives)** and relatively well for the target Metric, **Specificity (True Negatives / (True Negatives + False Positives)**
- Being that the primary objective for the business is to minimize False Positives (predicted to Renew but actually Defaulting/Non-Renewing), **a False Positive score of 2.5% is strong**
    - Although the model scored lower in Recall/Sensitivity (True Positives / (True Positives + False Negatives)), in this business scenario that simply equates to customers predicted to Not Renew but actually Renewing
        - **The overall False Negative count (22.5%) can be improved upon through Adjusted Probability Thresholds**
        - As long as this miss doesn't result in large amounts of resources wasted unnecessarily on targeting customers not truly at risk, it can be seen as a slight benefit to the business, with the key focus being targeting true at risk customers (likely only 6% of the population based on sample splits)

Pipeline: Adaptive Boost Classifier – Threshold of 0.48

```
Accuracy: 0.9043
Precision: 1.0
Recall: 0.885
F1 Score: 0.939
Specificity: 1.0
.................................
Roc_Auc_Score: 1.0

TP: 11764, FN: 1529, TN: 2678, FP: 0
```

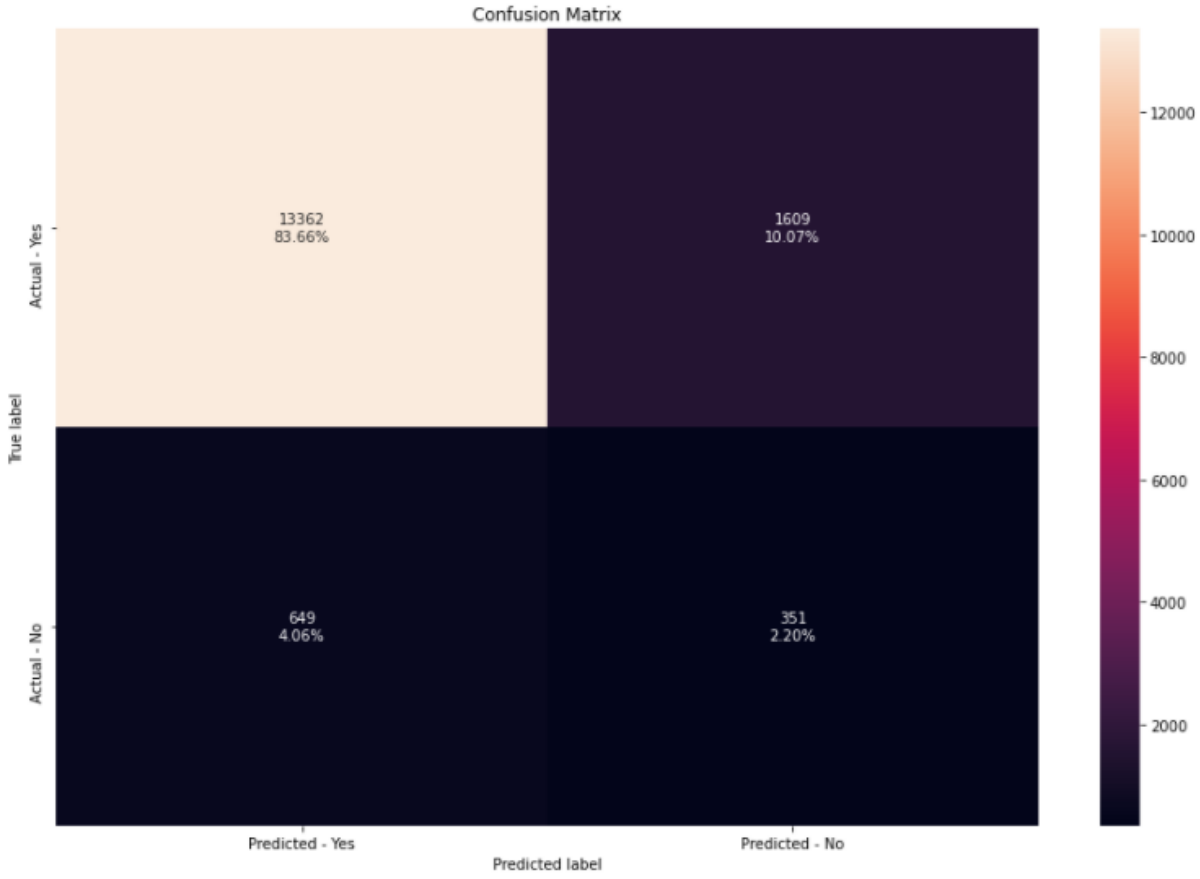


- **Lowering the threshold slightly down to 0.48** (vs. default of 0.5), after which point a customer is predicted to Renew, **lowers the False Negative Count from 22.5% to 9.6% and lowers the False Positive Count to 0%**
    - This results in perfect scores for Precision and Specificity, and substantially improves Recall (88.5%)

Pipeline: Gradient Boost Classifier

```
Accuracy: 0.8586
Precision: 0.9537
Recall: 0.8925
F1 Score: 0.9221
Specificity: 0.351
.................................
Roc_Auc_Score: 0.7628

TP: 13362, FN: 1609, TN: 351, FP: 649
```

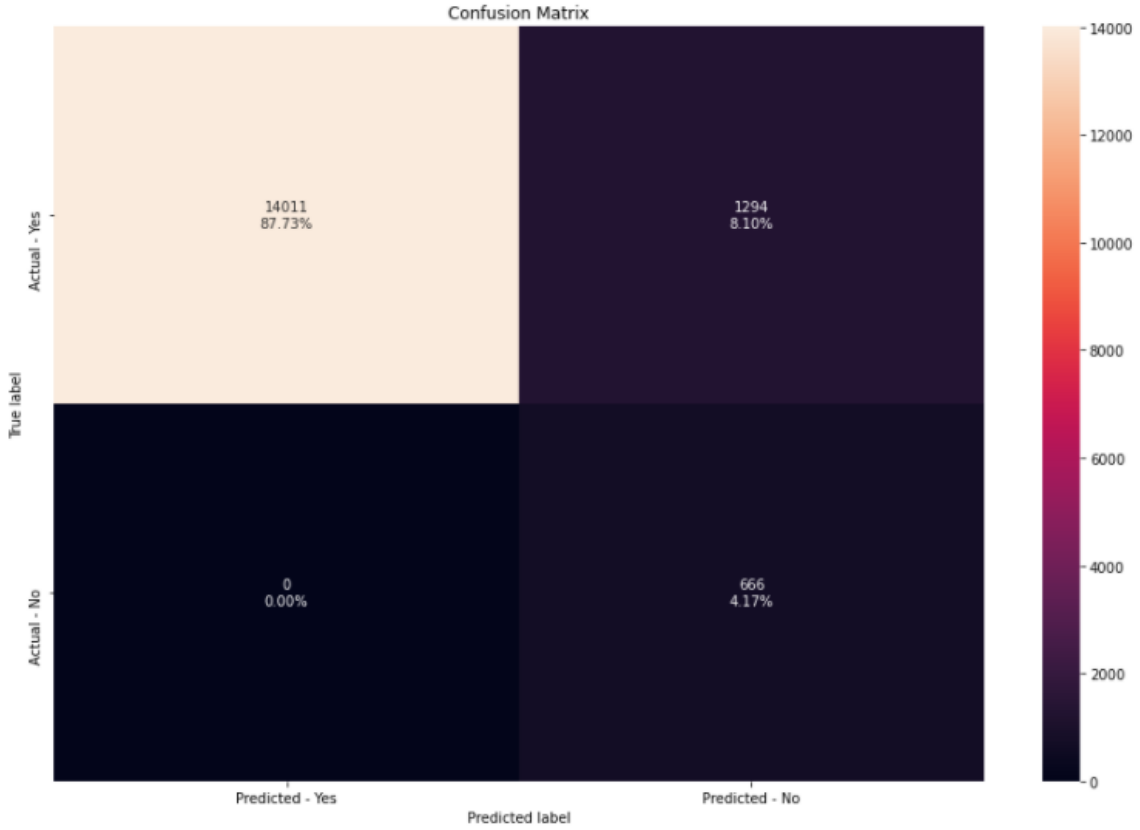


- The model scored very high for **Precision (True Positives / (True Positives + False Positives)** and poorly for the target Metric, **Specificity (True Negatives / (True Negatives + False Positives)**
- Being that the primary objective for the business is to minimize False Positives (predicted to Renew but actually Defaulting/Non-Renewing), **a False Positive score of 4.1% is decent**
    - Although the model scored lower in Recall/Sensitivity (True Positives / (True Positives + False Negatives)), in this business scenario that simply equates to customers predicted to Not Renew but actually Renewing
        - **The overall False Negative count (10.1%) may still be improved upon through Adjusted Probability Thresholds**
        - As long as this miss doesn't result in large amounts of resources wasted unnecessarily on targeting customers not truly at risk, it can be seen as a slight benefit to the business, with the key focus being targeting true at risk customers (likely only 6% of the population based on sample splits)

Pipeline: Gradient Boost Classifier – Threshold of 0.35

```
Accuracy: 0.919
Precision: 1.0
Recall: 0.9155
F1 Score: 0.9559
Specificity: 1.0
.................................
Roc_Auc_Score: 1.0

TP: 14011, FN: 1294, TN: 666, FP: 0
```



- **Lowering the threshold down to 0.35** (vs. default of 0.5), after which point a customer is predicted to Renew, **lowers the False Negative Count from 10.1% to 8.1% and lowers the False Positive Count to 0%**
    - This results in perfect scores for Precision and Specificity, and improves Recall (91.6%)

Best Model Selected

Depending on the final scenario chosen, there are 2 possible models worth considering as final for which to apply Test data to:

Scenario 1: Logistic Regression Pipeline

- Optimal Specificity score achieved (63%), with strong Precision and decent Recall scores
- Adjusted Probability Threshold of 40% (for Renewal) resulting in Perfect Precision/Specificity (0 False Positives) and Strong Recall (92%)

Standard Threshold of 0.5

- Accuracy: 0.76
- Precision: 0.97
- Recall: 0.77
- F1 Score: 0.86
- **Specificity: 0.63**

Adjusted Threshold of 0.4

- Accuracy: 0.93
- Precision: 1.0
- Recall: 0.92
- F1 Score: 0.96
- **Specificity: 1.0**

Scenario 2: Gradient Boost Classifier Pipeline

- Lowest Initial Specificity score achieved (35%), but with very stronger Recall and similar Precision scores
- Lower adjusted Probability Threshold of 35% (for Renewal) resulting in Perfect Precision/Specificity (0 False Positives) and Strong Recall (91.6%)

Standard Threshold of 0.5

- Accuracy: 0.86
- Precision: 0.95
- Recall: 0.89
- F1 Score: 0.92
- **Specificity: 0.35**

Adjusted Threshold of 0.35

- Accuracy: 0.92
- Precision: 1.0
- Recall: 0.92
- F1 Score: 0.96
- **Specificity: 1.0**

Decision: Logistic Regression Pipeline

Although having an initially lower Recall score, and requiring a slightly higher Probability Threshold for achieving similar results to the Gradient Boost Classifier, the **Logistic Regression model is better suited to the overarching business goal of addressing at risk customers most likely to default/non-renew their policies**.

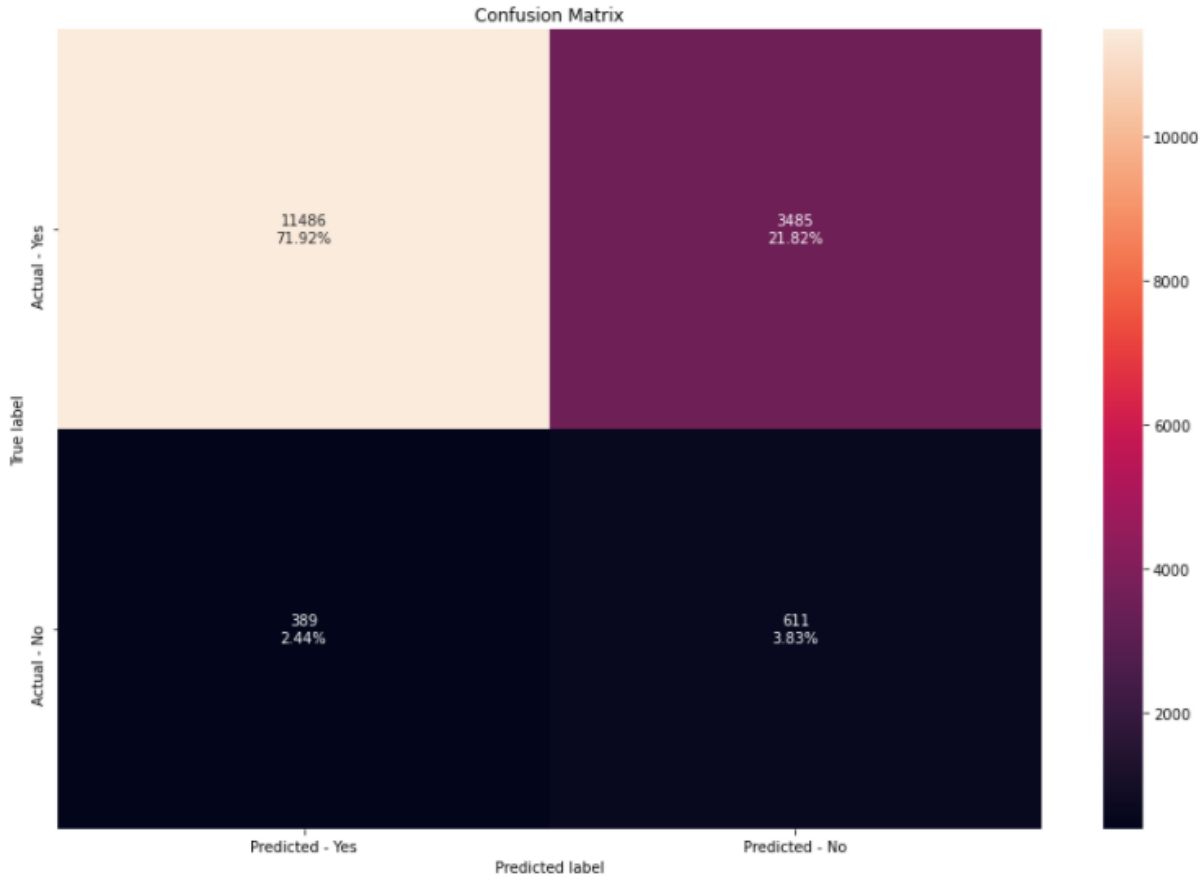
If costs of incorrect at-risk customer predictions (False Positives - Recall) also become a primary objective for the company, then **further Probability Threshold testing (0.4 or lower) can be adjusted to help reduce the False Positive counts**, which still accurately reducing False Negatives (incorrectly predicted Defaulters).

Confusion Matrix – Test Data

Pipeline: Logistic Regression

```
Accuracy: 0.7574
Precision: 0.9672
Recall: 0.7672
F1 Score: 0.8557
Specificity: 0.611
.................................
Roc_Auc_Score: 0.7573

TP: 11486, FN: 3485, TN: 611, FP: 389
```

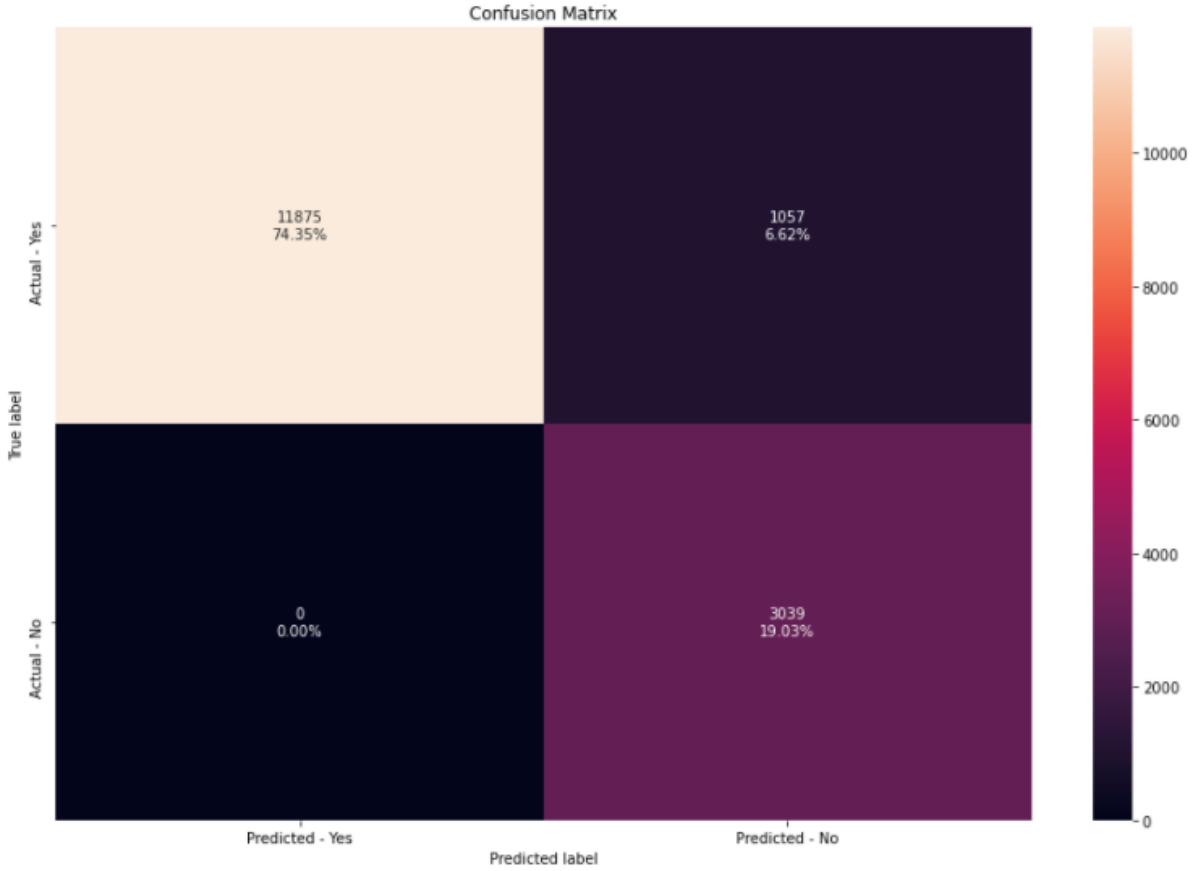


- The model scored very high for **Precision (True Positives / (True Positives + False Positives)** and relatively well for the target Metric, **Specificity (True Negatives / (True Negatives + False Positives)**
- Being that the primary objective for the business is to minimize False Positives (predicted to Renew but actually Defaulting/Non-Renewing), **a False Positive score of 2.4% is strong**
    - Although the model scored lower in Recall/Sensitivity (True Positives / (True Positives + False Negatives)), in this business scenario that simply equates to customers predicted to Not Renew but actually Renewing
        - **The overall False Negative count (21.8%) can be improved upon through Adjusted Probability Thresholds**
        - As long as this miss doesn't result in large amounts of resources wasted unnecessarily on targeting customers not truly at risk, it can be seen as a slight benefit to the business, with the key focus being targeting true at risk customers (likely only 6% of the population based on sample splits)

Pipeline: Logistic Regression – Threshold of 0.4

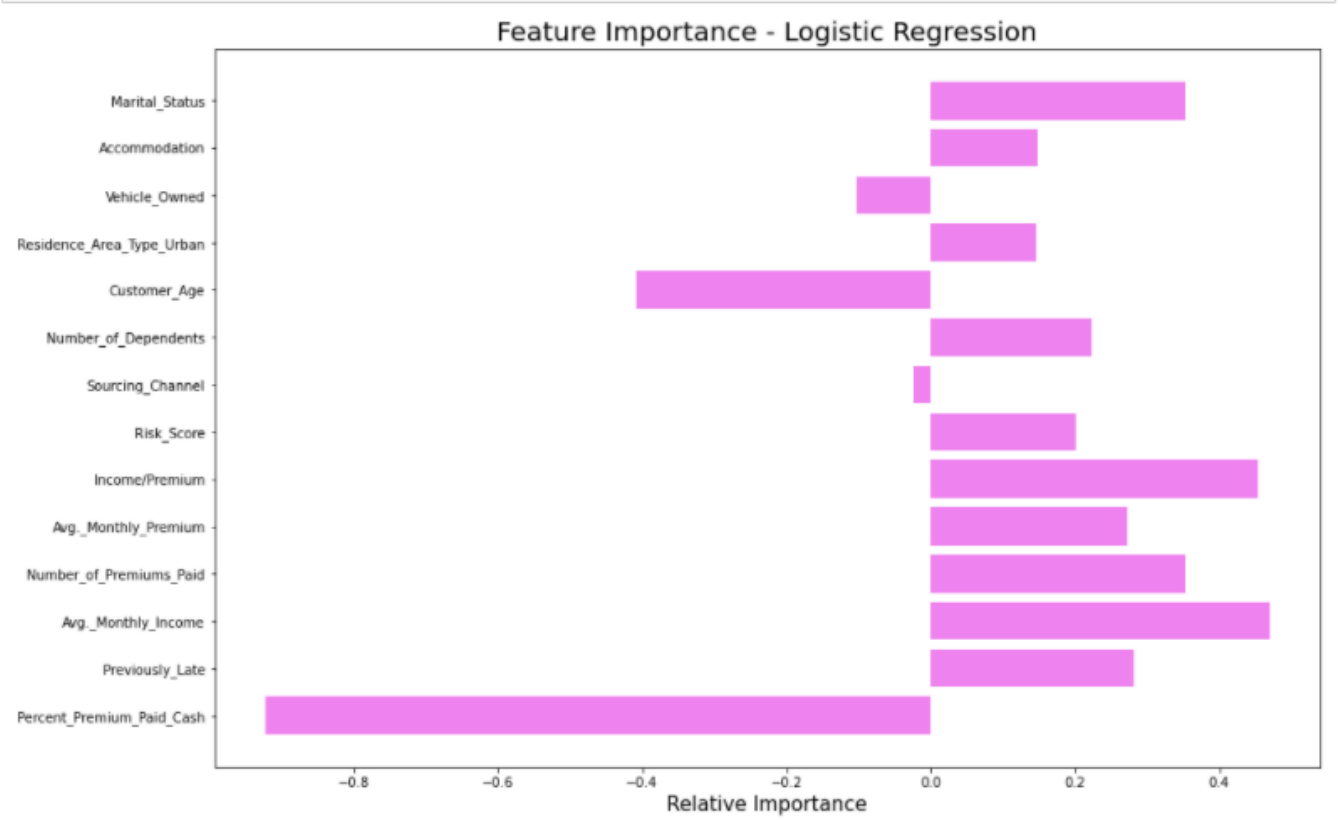
```
Accuracy: 0.9338
Precision: 1.0
Recall: 0.9183
F1 Score: 0.9574
Specificity: 1.0
.................................
Roc_Auc_Score: 1.0

TP: 11875, FN: 1057, TN: 3039, FP: 0
```



- **Lowering the threshold to 0.4** (vs. default of 0.5), after which point a customer is predicted to Renew, **lowers the False Negative Count from 21.8% to 6.6% and lowers the False Positive Count to 0%**
    - This results in perfect scores for Precision and Specificity, and substantially improves Recall (91.8%)

Key Feature Summary – Pipeline: Logistic Regression



As it relates to a customer's likelihood of Renewal (Majority Class-1), the top Features likely to decide the outcome are: **Less likely to Renew - Increasing Likelihood of Non-Renewal**

- Percent Premium Paid Cash (-0.92)
- Customer Age (-0.41)

More likely to Renew - Decreasing Likelihood of Non-Renewal

- Avg. Monthly Premium (0.47)
- Income/Premium (0.45)
- Number of Premiums Paid (0.35)
- Marital Status (0.35)

Business Insights & Recommendations

Scoring for Specificity

Although the models are performing well in regards to Precision, Recall, and Overall Accuracy, **special attention needs to be placed on better Specificity performance, particularly as it relates to identifying False and lowering the counts of False Positives (predicted to renew but actually not renewing)**.

Due to the imbalanced data issues, it is easy for a model to perform with 95% plus Precision simply due to the correctly predicting the majority class, however this is only half the battle – **the real value worth targeting their time/resources is in better False Positive targeting and customer identification**.

Key Customers to Target (based on Feature Importance)

In order to as efficiently and effectively target at-risk customers likely to default on their premiums, the company should pay specific attention to those customers who have paid a substantial amount of their premium with cash, possibly segmenting those customers into more targeted groupings (e.g. 50%, 75%, 90% or higher, etc.)

Particular focus should also be centered on customers of varying age groups for different targeted campaigns.

- Customers 40 years or younger will require more aggressive, incentivized marketing and communications as they show the greatest likelihood of Non-Renewal in general
    - This is largely due to the general assumption of a majority of members in this age group not needing to claim in the near future therefore questioning the need for continued premium payments
    - This age group is crucial for positive cash flows and every effort should be made to keep their accounts active
- Customers between 40 and 70 appear to the most consistent and likely to renew and maintain their policies in general
    - Periodic communications should be made in order to keep this base engaged and ensure their satisfaction
    - This age group is more likely to have claims during their tenure, so this somewhat boosts their likelihood of continued insurance renewals to maintain protection
- Customers 70 and older are less consistent (some canceling policies and others renewing into their final years) and harder to predict and categorize
    - Further sampling and research needs to be focused around understanding the needs and lifestyles of this customer segment to better cater to and market for their needs, while ensuring adequate claims coverage and policy protections