Insurance Premium

Renewal Propensity Analysis

Sven Meydell

Business Objective

- Insurance premiums paid, and renewed, 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
  - Worst case, **inadequate default prevention could negatively affect the entire insurance industry**
- The goal of this analysis is to <u>predict the probability that a customer will not renew their premium payment</u>
  - This will **help mitigate potential future losses, while maintaining healthy cash-flows**
  - Correct implementation with improve the overall safety structure of the insurance/reinsurance partnership that keeps everything afloat **(premiums in/claims out)**

Data Provided

- **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**: Annual income of the customer
- **Premium**: Annual premium paid by 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

Note on Target Variable

- We wish to predict which customers with upcoming premiums are likely to default (not renew) their policy, regardless of whether/not they were previously late - **The Renewal column is therefore the target variable**

Data Modifications

Data Preprocessing

- **ID**: Kept intact initially (utilized post model building analysis)
- **Age in Days**: Converted to Age in years – divided by 365 days
- **Income**: Removed and replaced by:
  - **Avg. Monthly Income**
  - **Income/Premium**
- **Premium**: Removed and replaced by:
  - **Avg. Monthly Premium**
  - **Income/Premium**
- **All Counts of Months Late (3-12+ Months)**: Converted to binary Previously Late column (No/0 or Yes/1)
- **Premium Renewal:** Renamed to Renewed Policy and values transformed to string (No or Yes) for initial Exploratory Data Analysis

Outlier Treatment

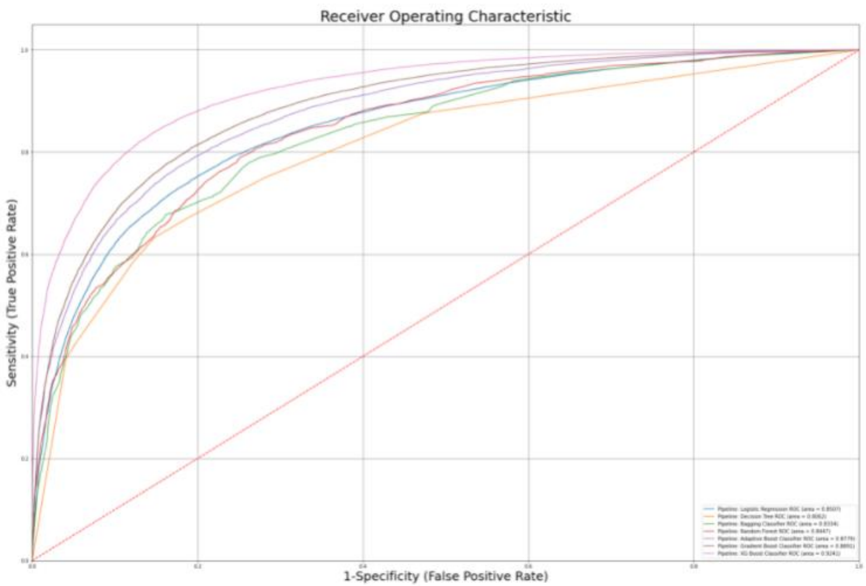
All outliers capped at IQR +/- 1.5x – Affected Variables are:

- Customer Age, Risk Score
- Number of Premiums Paid, Avg. Monthly Premium, Avg. Monthly Income, Income/Premium

Model Encoding

- **Renewed Policy (Target):** Manually converted back to Numeric (No/0 or Yes/1)
- **Sourcing Channel:** Label Encoded (Ordinal) for Channels A-E as numeric values 0-4
- **Residence Area Type:** One-Hot Encoded with first column (Rural) dropped

Top 3 Models – ROC Curves & CV Scores

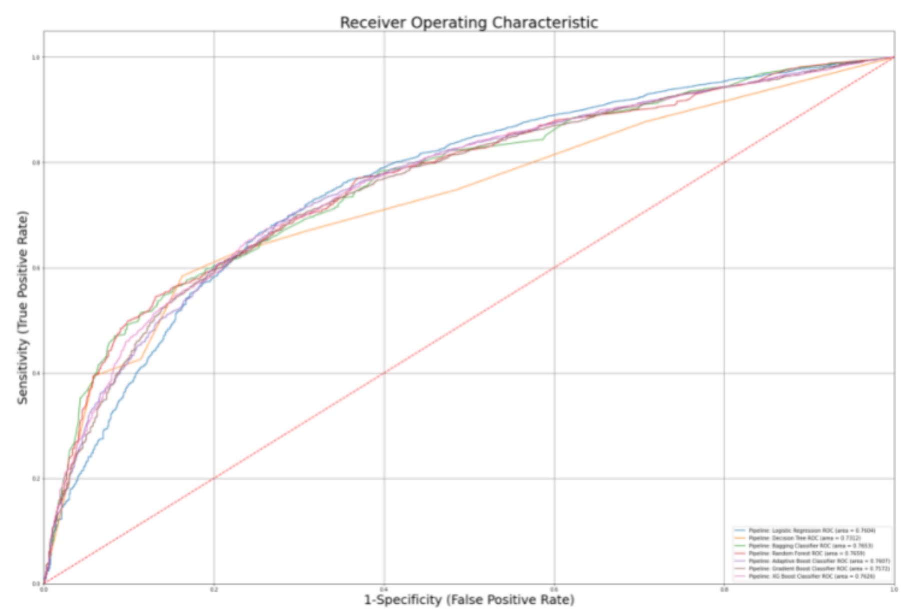


Cross Validation - ROC_AUC Score
(Train Data):

Pipeline: Logistic Regression: 85.1%
Pipeline: Decision Tree: 80.5%
Pipeline: Bagging Classifier: 83.6%
Pipeline: Random Forest: 84.2%
Pipeline: Adaptive Boost Classifier: 87.5%
Pipeline: Gradient Boost Classifier: 88.7%
Pipeline: XG Boost Classifier: 91.1%

Cross Validation - ROC_AUC Score
(Validation Data):

Pipeline: Logistic Regression: 82.4%
Pipeline: Decision Tree: 80.3%
Pipeline: Bagging Classifier: 82.2%
Pipeline: Random Forest: 82.2%
Pipeline: Adaptive Boost Classifier: 81.9%
Pipeline: Gradient Boost Classifier: 82.5%
Pipeline: XG Boost Classifier: 81.4%



Best Model Selection

Depending on the final scenario chosen, there are 2 possible models worth considering as final for the Test Dataset

- **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%)

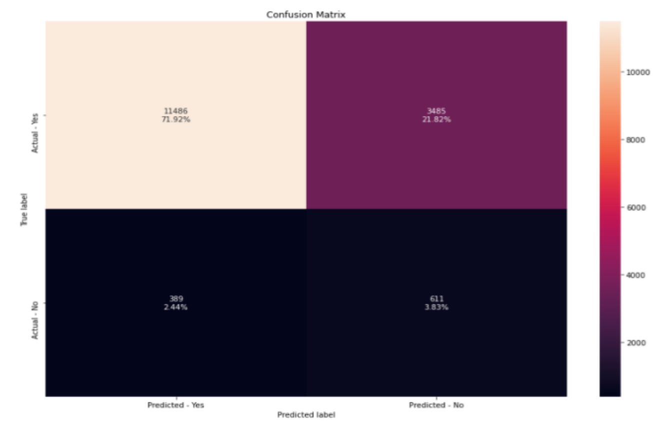
- **Scenario 2: Gradient Boost Classifier Pipeline**
  - Lowest Initial Specificity score achieved (35%), but with **very strong Recall and similar Precision scores**
  - **Lower adjusted Probability Threshold of 35%** (for Renewal) resulting in Perfect Precision/Specificity (0 False Positives) and Strong Recall (91.6%)

Final Decision: Logistic Regression Pipeline

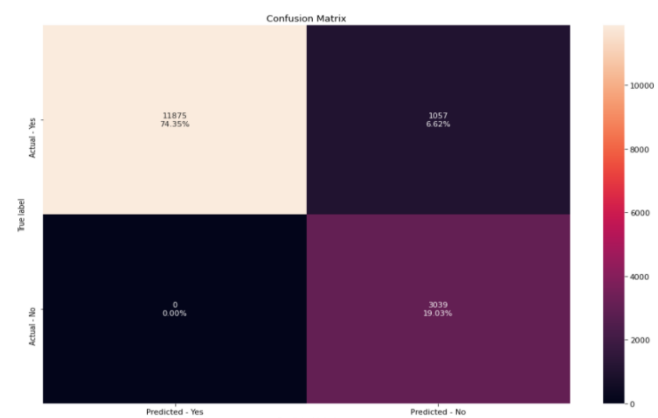
- The model is **better suited to the overarching business goal of addressing at-risk customers** most likely to default/non-renew their policies (Specificity Score)

- When the cost of incorrect at-risk customer predictions (False Positives - Recall) also become a primary objective
  - **New Probability Thresholds testing (0.4 or lower)** can be adjusted to help reduce the False Positive counts
  - This can simultaneously reduce False Negatives – improving Recall score

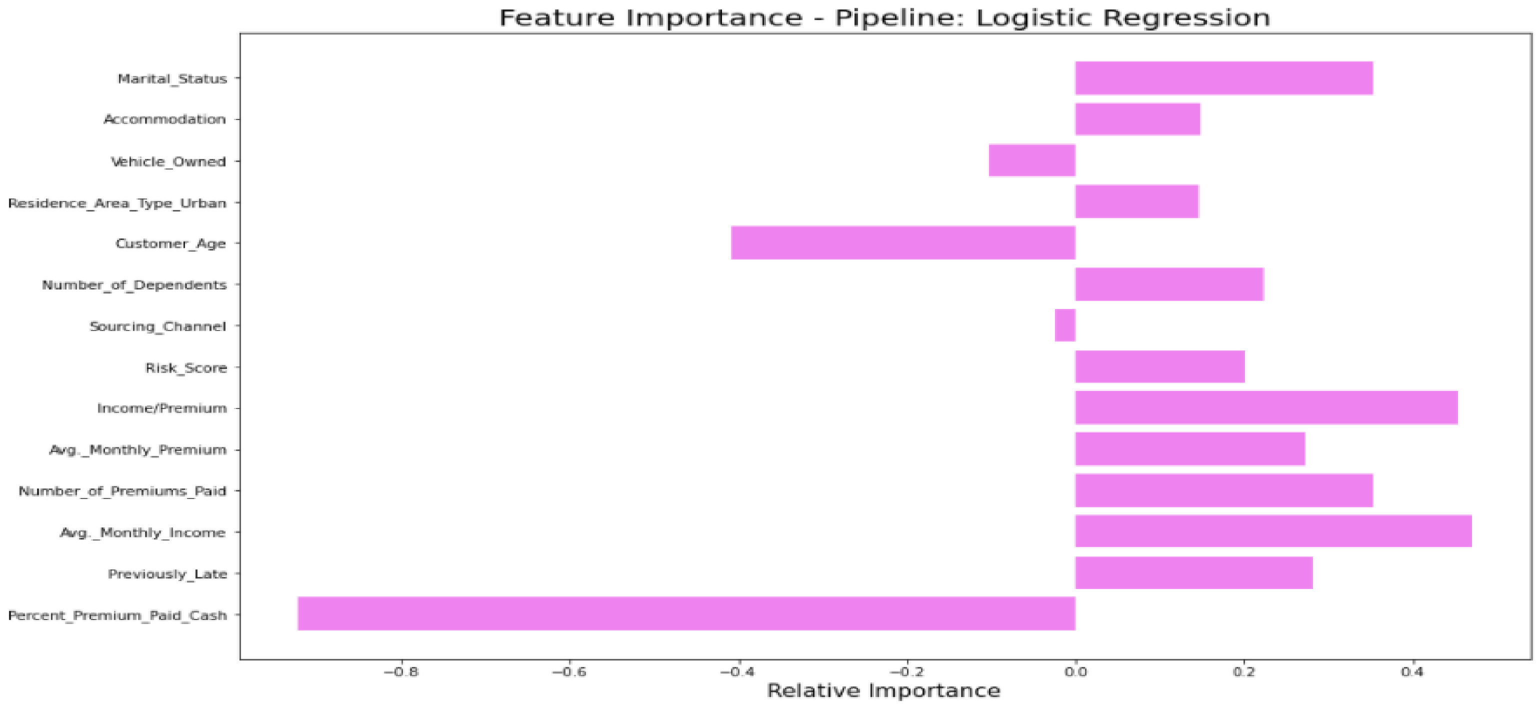
Logistic Regression - Regular



Logistic Regression – 0.40 Threshold



Key Features (Model Coefficients)



Feature Summary

As it relates to a customers likelihood of Renewal (Majority Class-1), the top Features likely to decide the outcome are:

Less likely to Renew - Increasing Likelihood of Non-Renewal

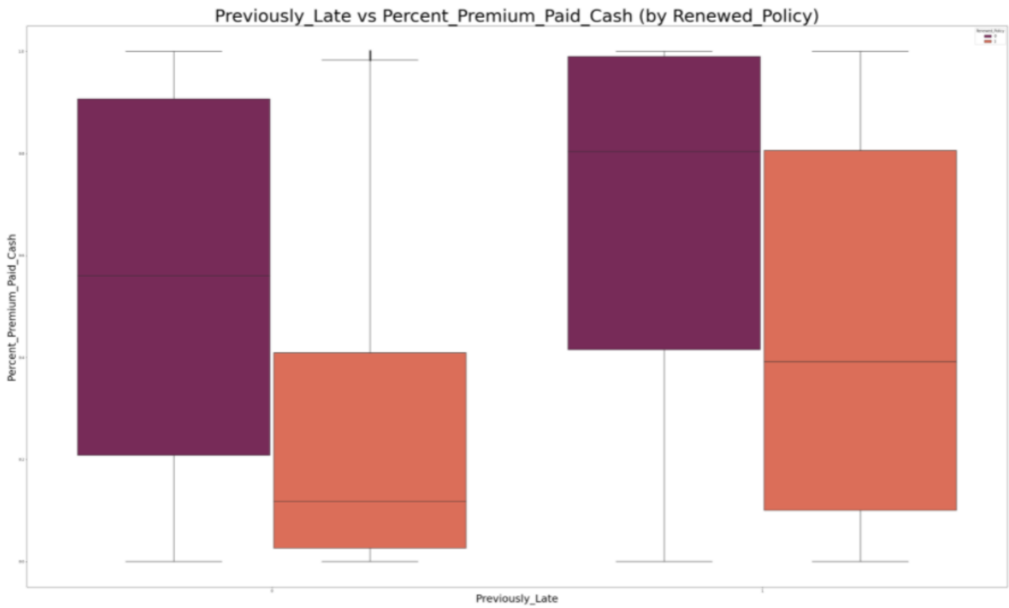
- Percent Premium Paid Cash (-0.92)
- Customer Age (-0.41)

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)

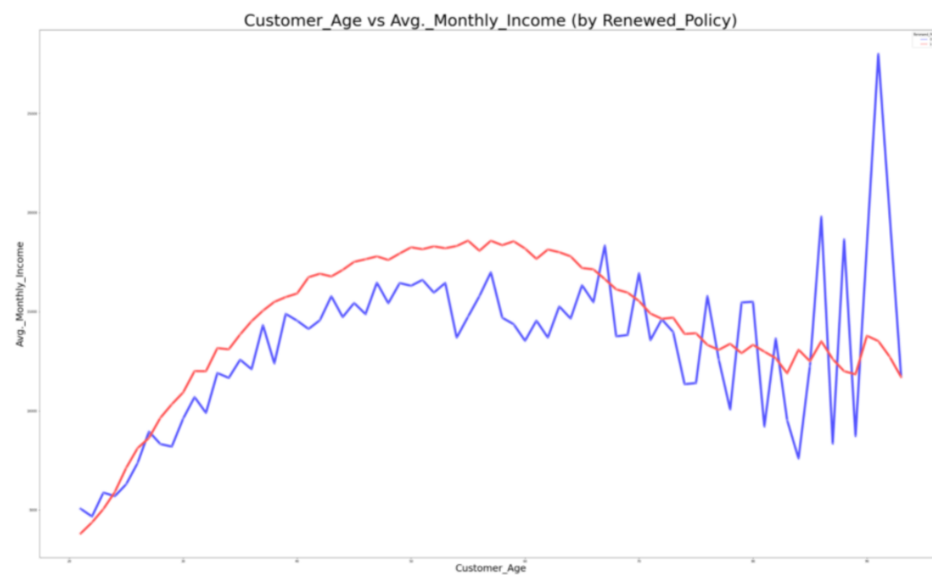
* Above scores indicate how much the **mean of the dependent variable (target) changes given a one-unit shift in the independent variable** while holding other variables in the model constant

EDA – Interaction Analysis (w. Target)



- In general, customers that have **paid higher portions of their policy with cash are far more likely** to:
    - Have been, or will be, late on their accounts,
    - Potentially default (non-renewing) their policies

- **Average monthly income increases with age**, up to a point from around 50 to 60
    - Incomes then drop as is somewhat expected as individuals transition out of full-time employment
- There are large spikes in average monthly income, as well as non-renewals, for some customers in their **mid-80s through mid-90s**
    - This could be indicative of customers **cashing out various (final) retirement products**



Recommendations

Key Customers to Target (based on Feature Importance)

- The company should pay specific attention to those customers who have paid a substantial amount of their premium with cash
  - Possible segmentation of customers into **more targeted groupings (e.g. 30%, 50%, 70% Paid Cash, etc.)**

- Particular focus should also be centered on customers of varying age groups for different targeted campaigns
  - Customers 40 years or younger - **greatest likelihood of Non-Renewal in general**
    - More aggressive, incentivized marketing/communications
  - Customers between 40 and 70 appear - **the most consistent and likely to renew**
    - Oftentimes most likely to claim – constantly review/maintain risk scores
  - Customers 70 and older - **less consistent and harder to predict/categorize**
    - Sometimes canceling policies and others renewing into their final years
    - Harder to predict and categorize overall
    - Consider dissecting and analyzing further as subgroups (Renewing vs. Non-Renewing)

Additional Insights

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**
  - Due to the imbalanced data issues, it is easy for a model to perform with 94% plus Precision simply due to the correctly predicting only the majority class
  - Specificity therefore is crucial for specifically identifying and **lowering the counts of False Positives** (At-Risk Customers predicted to renew but actually not renewing their policies)