**Titulo**

**Miguel Coto García, carnet**

**Natalia Díaz Ramírez, B12251**

**Universidad de Costa Rica**

**Programa de Posgrado en Estadística**

**SP1626, Estadística Bayesiana**

**Introducción**

El sector asegurador es altamente competitivo donde la diversidad en el mercado ha originado que las empresas se enfrenten al abandono de los clientes. Según Gutiérrez et al (2017) “cada vez es mayor el número de clientes que son conscientes de que existen otras alternativas de productos o servicios en el mercado y que no tienen inconveniente en comparar ofertas y en buscar otras opciones de compra, así como, llegado el momento, abandonar a su empresa y cambiar de marca”. (pág.77)

La caída de la cartera es de importancia para el sector asegurador ya que existe el riesgo de la pérdida de clientes debido a la cancelación de pólizas, por lo que surge la necesidad de retener y fidelizar a los clientes para garantizar la rentabilidad de la empresa. La pérdida de la lealtad podría servir como indicador para predecir el posible abandono, la cual puede ir ligada a una reducción del compromiso afectivo del cliente con la empresa o a una disminución de su vinculación en términos de frecuencia de compra de productos o de cancelación de servicios contratados. (Gutiérrez et al, 2017, pág.78)

A nivel económico de la empresa aseguradora Guillén et al (2015) cita que, para una cartera de pólizas, los gastos de emisión suelen superar a la prima cobrada el primer año. Es decir, que, si una póliza deja de ser pagada dentro de su primer año, genera pérdidas para el asegurador. (Pág. 3)

Respecto al riesgo de la caída de la cartera Arce et al (2011) menciona que “en un competitivo

mercado asegurador donde cada día toman relevancia temas como la guerra de precios, accesibilidad a múltiples cotizaciones, así como la constante innovación en el desarrollo de productos, surge la necesidad de retener y fidelizar a los clientes”. (Pág.58)

La retención de clientes es una estrategia para garantizar rentabilidad y éxito en una compañía, por lo que es necesario tomar acciones preventivas que eviten que los clientes decidan abandonar los servicios brindados por la compañía y promover la fidelidad de dichos clientes con la entidad.

El cliente es parte fundamental de una empresa por lo que comprender el tipo de cliente que esta propenso a cancelar las pólizas de seguro permite desarrollar estrategias para retener y fidelizar a esos clientes.

**Método**

El objetivo del presente estudio es identificar los factores que influyen en la propensión de cancelación de pólizas de una aseguradora.

Los datos son de una aseguradora, correspondiente a pólizas vendidas entre enero de 2017 hasta abril de 2018. La base de datos tiene 20192 pólizas con una antigüedad de tres meses o más. Las variables para el análisis son las siguientes:

*Variable dependiente:*

* Estado: Corresponde al estado de la póliza cancelada (1) o vigente (0).

*Variables independientes:*

Variables numéricas

* Antigüedad: antigüedad de la póliza en meses.
* Edad: Edad del cliente.
* Número de hijos: Cantidad de hijos que tiene el cliente.

Variables categóricas

* Producto: Tipo de póliza (por confidencialidad se clasificaron en Producto A, B, C, D, E)
* Moneda: Tipo de moneda de la prima de la póliza (colones o dólares)
* Plan: Tipo de pago de la prima (mensual o anual)
* Medio de pago por emisor: corresponde al tipo de emisor de la tarjeta con que se realiza el pago de la póliza (por confidencialidad se clasificaron en Emisor A y Emisor B)
* Medio de pago por plan: corresponde al tipo de plan de la tarjeta con que se realiza el pago de la póliza (por confidencialidad se clasificaron en Plan A, B, C, D, E)
* Sexo: Sexo del cliente.

Los datos son desbalanceados, ya que hay un 30% de pólizas canceladas y un 70% de pólizas vigentes.

Se realizó una selección de variables mediante el método bayesiano de Metrópolis Hastings utilizando el paquete MCMClogit para un modelo de regresión logística. Se compararon cuatro modelos con cantidades de variables diferentes para cada uno y se corrió una cadena de 30000 iteraciones, de las cuales fueron descartadas las 1000 primeras. Se utilizó el factor de bayes para determinar el modelo más probable. Los modelos utilizados siguen las siguientes distribuciones, asumiendo una distribución a priori normal multivariada para los coeficientes:

Una vez obtenidas las variables más importantes se realizó un análisis bayesiano vía muestreo de Gibbs utilizando el paquete MCMCglmm del software R para una regresión logística. Donde se compararon seis modelos variando las distribuciones a priori y utilizando el DIC como criterio de comparación de modelos.

Distribuciones:

* Priori plana no informativa
* Priori plana no informativa para los betas y gamma para el error
* Priori plana informativa para betas
* Priori de Gelman para betas
* Priori plana informativa para betas y errores fijados en 1

En este caso se utilizó el 75% (15144) de los datos como base para entrenar el modelo y un 25% (5048) de los datos para validar el modelo. A partir de las probabilidades predichas se utilizó un punto de corte de 0.3, donde probabilidades mayores a este son clasificadas como canceladas. Las medidas de precisión utilizadas fueron la precisión global, que indica que porcentaje del total se clasificó correctamente tomando en cuenta ambas categorías y la precisión positiva, que hace referencia al porcentaje de clasificaciones de canceladas realizadas por el modelo correctamente.

Para verificar la convergencia de los modelos realizados se utilizó el gráfico de autocorrelación, el gráfico de Geweke y la prueba de Heidelberger y Welch.

**Resultados**

**Selección de variables**

Se obtuvo que el modelo más probable es el modelo que tiene como variables predictoras el producto, el medio de pago por plan, la antigüedad de la póliza y la edad del cliente, según la tabla 1 este corresponde al modelo 4.

|  |  |  |  |
| --- | --- | --- | --- |
| **Tabla 1. Comparación de modelos para selección de variables** | | | |
| **Modelo** | **Variables** | **Convergencia** | **Probabilidad Posterior Factor Bayes** |
| **Modelo 1** Completo | Producto+Moneda+Plan+MedioPagoEmisor+MedioPagoPlan+Antiguedad+edad+sexo+Numero\_Hijos | No | 0.0000 |
| **Modelo 2** | Producto+Moneda+Plan+MedioPagoEmisor+MedioPagoPlan+Antiguedad+edad | No | 0.0000 |
| **Modelo 3** | Producto+MedioPagoEmisor+MedioPagoPlan+Antiguedad+edad | No | 0.0015 |
| **Modelo 4** | Producto+MedioPagoPlan+Antiguedad+edad | Si | 0.9985 |

Los resultados de las pruebas de convergencia del modelo 4 se puede ver en los anexos 1, 2 y 3, los cuales nos indican que se alcanza convergencia.

En la tabla 2, se muestra, a partir de los intervalos de credibilidad que las 4 variables seleccionadas son relevantes en la propensión de cancelación de pólizas.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tabla 2. Coeficientes e intervalo de credibilidad del Modelo 4** | | | | |
| **Variables** | **Media Posterior** | **Desv.Estándar** | **2.5%** | **97.5%** |
| (Intercept) | -2.078 | 0.173 | -2.429 | -1.744 |
| ProductoB | 0.360 | 0.040 | 0.279 | 0.439 |
| ProductoC | 0.375 | 0.055 | 0.272 | 0.484 |
| ProductoD | 0.885 | 0.288 | 0.318 | 1.406 |
| ProductoE | 0.916 | 0.479 | -0.076 | 1.806 |
| MedioPagoPlanPlanB | 0.314 | 0.162 | 0.012 | 0.649 |
| MedioPagoPlanPlanC | 0.438 | 0.159 | 0.143 | 0.770 |
| MedioPagoPlanPlanD | 0.642 | 0.161 | 0.351 | 0.975 |
| MedioPagoPlanPlanE | 0.775 | 0.160 | 0.474 | 1.092 |
| Antiguedad | 0.132 | 0.005 | 0.123 | 0.142 |
| edad | -0.017 | 0.001 | -0.019 | -0.014 |

**Modelo co**n **MCMCglmm**

Se compararon los primeros cinco modelos de la tabla 3 variando la distribución a priori. Se obtuvo que el modelo que obtuvo un menor DIC fue el modelo 2, sin embargo, al evaluar la convergencia dicho modelo no converge. El modelo 5 obtuvo mejoras en los gráficos evaluados para la convergencia, aunque no se alcanza convergencia, por lo que se decidió aumentar la cantidad de iteraciones de este modelo, el cual está representado en el modelo 6. El modelo 6, asume convergencia con un total de 30000 iteraciones por lo que este fue el modelo seleccionado, el detalle de la convergencia de dicho modelo se puede consultar en los anexos 4, 5 y 6.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tabla 3. Parámetros usados y resultados de los modelos** | | | | | | |
| **Modelo** | **A priori** | **Parámetros de modelos** | | | **DIC** | **Convergencia** |
| **nitt** | **burnin** | **thin** |
| Modelo1 | Priori plana no informativa | 13000 | 3000 | 10 | 17500.64 | No |
| Modelo2 | priori plana no informativa para los betas y gamma para el error | 13000 | 3000 | 10 | 16983.1 | No |
| Modelo3 | priori plana informativa para betas | 13000 | 3000 | 10 | 17154.68 | No |
| Modelo4 | priori de Gelman para betas | 13000 | 3000 | 10 | 17548.53 | No |
| Modelo5 | priori plana informativa para betas y errores fijados en 1 | 13000 | 3000 | 10 | 17464.59 | No (Pero Gráficos Mejoran) |
| Modelo6 | priori plana informativa para betas y errores fijados en 1 | 30000 | 3000 | 20 | 17462.89 | Si |

Además, en la tabla 4 se observan las medidas de precisión de las predicciones realizadas con cada modelo, las cuales son muy similares lo que indica que utilizar aprioris diferentes no afecta las estimaciones de las probabilidades predichas.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tabla 4. Medidas de Precisión de los modelos** | | | | |
|
| **Modelo** | **A priori** | **Precisión Global** | **Precisión Positiva** | **Precisión Negativa** |
|
| Modelo1 | priori planas no informativas | 0.621 | 0.631 | 0.617 |
| Modelo2 | priori plana no informativa para los betas y gamma para el error | 0.620 | 0.632 | 0.615 |
| Modelo3 | priori plana informativa para betas | 0.620 | 0.633 | 0.614 |
| Modelo4 | priori de Gelman para betas | 0.622 | 0.629 | 0.619 |
| Modelo5 | priori plana informativa para betas y errores fijados en 1 | 0.621 | 0.631 | 0.616 |
| Modelo6 | priori plana informativa para betas y errores fijados en 1 | 0.621 | 0.629 | 0.617 |

En cuanto al modelo seleccionado en el anexo 7 se encuentran los gráficos de la traza y la distribución posterior del modelo.

En la tabla 5, se presentan las estimaciones de los coeficientes para el modelo seleccionado, donde se observa que los coeficientes del Producto E y el medio de pago con plan B no son relevantes en la propensión de cancelación de seguros, ya que el intervalo de credibilidad incluye al cero.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tabla 5. Estimación de los coeficientes con el modelo seleccionado** | | | | | | |
| **Coeficientes** | **Media Posterior** | **l-95% CI** | **u-95% CI** | **OR** | **l-95% CI** | **u-95% CI** |
| (Intercept) | -2.34 | -2.84 | -1.88 | 0.10 | 0.06 | 0.15 |
| ProductoB | 0.41 | 0.31 | 0.52 | 1.51 | 1.37 | 1.69 |
| ProductoC | 0.43 | 0.30 | 0.60 | 1.54 | 1.35 | 1.82 |
| ProductoD | 0.96 | 0.29 | 1.68 | 2.62 | 1.34 | 5.35 |
| ***ProductoE*** | ***0.62*** | ***-0.66*** | ***2.04*** | ***1.87*** | ***0.52*** | ***7.73*** |
| ***MedioPagoPlanPlanB*** | ***0.32*** | ***-0.08*** | ***0.76*** | ***1.38*** | ***0.92*** | ***2.14*** |
| MedioPagoPlanPlanC | 0.44 | 0.03 | 0.85 | 1.55 | 1.03 | 2.34 |
| MedioPagoPlanPlanD | 0.66 | 0.26 | 1.09 | 1.94 | 1.30 | 2.97 |
| MedioPagoPlanPlanE | 0.86 | 0.46 | 1.29 | 2.37 | 1.58 | 3.64 |
| Antiguedad | 0.15 | 0.14 | 0.16 | 1.16 | 1.15 | 1.18 |
| edad | -0.02 | -0.02 | -0.02 | 0.98 | 0.98 | 0.98 |

En la tabla 6 se interpretan los odds ratio estimados en el modelo seleccionado.

|  |  |
| --- | --- |
| **Tabla 6. Interpretación de las estimaciones de los odds ratio** | |
| **Variable** | **Interpretación** |
| Producto | Los odds de cancelación de las pólizas en el producto B son entre 37% y 69% mayor que los odds de cancelación de las pólizas en el producto A, manteniendo las demás variables constantes.  Los odds de cancelación de las pólizas en el producto C son entre 35% y 82% mayor que los odds de cancelación de las pólizas en el producto A, manteniendo las demás variables constantes.  Los odds de cancelación de las pólizas en el producto D son entre 34% mayor y 5 veces los odds de cancelación de las pólizas en el producto A, manteniendo las demás variables constantes. |
| Medio de pago por plan | Los odds de cancelación de las pólizas con un medio de pago por el plan C son entre 3% mayor y 2 veces los odds de cancelación de las pólizas con un medio de pago por plan A, manteniendo las demás variables constantes.  Los odds de cancelación de las pólizas con un medio de pago por el plan D son entre 30% mayor y 3 veces los odds de cancelación de las pólizas con un medio de pago por plan A, manteniendo las demás variables constantes.  Los odds de cancelación de las pólizas con un medio de pago por el plan E son entre 58% mayor y 3.6 veces los odds de cancelación de las pólizas con un medio de pago por plan A, manteniendo las demás variables constantes. |
| Antigüedad | Los odds de cancelación de las pólizas aumenta entre 15% y 18% por cada mes de antigüedad, manteniendo las demás variables constantes. |
| Edad | Los odds de cancelación de las pólizas disminuye un 2% por cada año de edad, manteniendo las demás variables constantes. |

En cuanto a la clasificación del modelo, se muestra la matriz de confusión y las medidas de precisión del modelo obtenido en la base de testeo.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tabla 7**  **Matriz de confusión con base de testeo** | | | | |
|  |  | **Predicho** | |  |
|  |  | 0 | 1 |  |
| **Real** | 0 | 2145 | 1330 | **3475** |
|  | 1 | 583 | 990 | **1573** |
|  |  | **2728** | **2320** | **5048** |

En la fase de validación del modelo se obtuvo una precisión global del 62% y una precisión positiva del 63%.

**Discusión**

Agregar más bibliografía

Entre los principales resultados se obtuvo que los factores que influyen en la cancelación de seguros, de la aseguradora estudiada, son el producto o tipo de seguro, el medio de pago por plan, la antigüedad de la póliza y la edad del cliente.

Además, se logró construir un modelo predictivo el cual logró clasificar correctamente un 63% de las pólizas canceladas. Dicho modelo presenta medidas de ajuste aceptables, a pesar de haber trabajado con un caso desbalanceado al predominar una categoría sobre otra, en este caso las pólizas vigentes presentan una mayor proporción que las canceladas.

Se identificaron factores asociados directamente a la póliza como la antigüedad y el tipo de seguro o producto. En este caso la propensión de cancelación es menor en el producto A que en los otros tipos de seguro. Además, conforme aumenta la antigüedad mayor es la propensión de cancelación, en los primeros años de vida de las pólizas es donde se presenta el riesgo de cancelación por lo que el resultado obtenido en la presente investigación tiene sentido, ya que las pólizas tienen una antigüedad entre 3 meses y 15.7 meses.

En cuanto al tipo de producto en el modelo realizado por Gutiérrez et al (2017) también obtuvieron que el tipo de seguro es una variable que tiene un efecto significativo en la cancelación de los seguros. Así mismo, Guillén et al (2015) obtuvo que el tipo de póliza incide sobre la probabilidad de cancelar principalmente las pólizas de salud.

El medio de pago por plan A tiene una menor propensión de cancelación que los otros medios de pago por plan, esto se puede explicar debido a que los clientes del tipo de plan A son de un perfil alto con una estabilidad económica sólida, cuyos ingresos tienden a ser altos.

Existen factores ligados al cliente como la edad. Entre los resultados se obtuvo que la propensión de cancelación disminuye por cada año de edad del cliente, lo cual concuerda con los resultados del estudio obtenido por Guillen et al (2017) el cual fue realizado con estadística clásica y una regresión logística.

**Referencias**

Arce, E., Mejía, M. (2011) Aplicación de un modelo de evaluación crediticia para reducir el riesgo en la cartera de clientes de una compañía aseguradora. *Industrial Data*: 14(2), pp. 59-66.

Guillén, M., Pérez, A. (2015)Riesgo de negocio ante asegurados con múltiples contratos. Fundación MAPFRE.

Gutiérrez, M., Segovia, M., Ramos M. (2017) Análisis del Riesgo de Caída de Cartera en Seguros: Metodologías de “Inteligencia Artificial” vs “Modelos Lineales Generalizados”. *Economía Informa*: 407, pp. 56-86.

**Anexo**

**Anexo 1**

|  |
| --- |
| **Prueba de Heidelberger y Welch del modelo final en selección de variables** |
|  |

**Anexo 2**

|  |
| --- |
| **Gráficos de autocorrelación del modelo final en selección de variables** |
|  |

**Anexo 3**

|  |  |
| --- | --- |
| **Gráficos de Geweke del modelo final en selección de variables** | |
|  |  |

**Anexo 4**

|  |
| --- |
| **Prueba de Heidelberger y Welch del modelo final** |
|  |

**Anexo 5**

|  |
| --- |
| **Gráficos de autocorrelación del modelo final** |
|  |

**Anexo 6**

|  |  |
| --- | --- |
| **Gráficos de Geweke del modelo final** | |
|  |  |

**Anexo 7**

|  |  |
| --- | --- |
| **Traza y distribuciones posteriores del modelo final** | |
|  |  |
|  |  |