**Titulo**

**Miguel Coto García,**

**Natalia Díaz Ramírez, B12251**

**Universidad de Costa Rica**

**Programa de Posgrado en Estadística**

**SP1626, Estadística Bayesiana**

**Introducción**

**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.

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 planas no informativas
* 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.

Para verificar la convergencia de los modelos realizados se utilizó el gráfico de autocorrelación, 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.estandar** | **2.50%** | **97.50%** |
| (Intercept) | -2.07822 | 0.173042 | -2.42938 | -1.74418 |
| ProductoB | 0.36002 | 0.040265 | 0.27973 | 0.43848 |
| ProductoC | 0.37528 | 0.055216 | 0.27166 | 0.48386 |
| ProductoD | 0.88532 | 0.288422 | 0.3176 | 1.40577 |
| ProductoE | 0.91593 | 0.479467 | -0.07611 | 1.80619 |
| MedioPagoPlanPlanB | 0.31407 | 0.162375 | 0.01219 | 0.64937 |
| MedioPagoPlanPlanC | 0.43765 | 0.159531 | 0.1431 | 0.77052 |
| MedioPagoPlanPlanD | 0.6416 | 0.160593 | 0.35096 | 0.97491 |
| MedioPagoPlanPlanE | 0.775 | 0.160167 | 0.47412 | 1.09242 |
| Antiguedad | 0.13227 | 0.00482 | 0.12284 | 0.14173 |
| edad | -0.01667 | 0.001341 | -0.01927 | -0.01403 |

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

**Discusión**

**Referencias**

**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** | |
|  |  |