Propuesta de Tesis

Franco Tralice

Objetivo

Desarrollar un modelo de predicción de bajas de clientes de seguros (churn analysis). Para esto se utilizarán técnicas de machine learning. Con estas predicciones, se podrán generar campañas para aumentar la retención de los clientes.

Aplicación

El modelo será utilizado para una empresa de Seguros de Argentina que proveeyó de los datos. Se podrá actualizar a medida que pasa el tiempo para poder realizar una comparación con algún grupo de control para ver si efectivamente se están recuperando clientes que se hubieran dado de baja.

Problema

El Customer Churn es el KPI que mide la cantidad de clientes que dejan de utilizar un producto o servicio. Para la mayoría de los bancos y compañías de seguros, el churn de seguros es un tema muy importante, ya que el costo de incorporar un cliente nuevo o el de volver a incorporar uno que ya se había dado de baja es mucho mayor al de retener uno ya existente.
En Argentina la venta de seguros no es muy buena, y va cayendo en los últimos años debido a diversas razones.

Modelo

El modelo tratará de exponer una probabilidad de baja de cada uno de los usuarios. Se probarán distintos algoritmos para poder tener una mayor precisión. Se tratará de investigar si es mejor modelar dividiendo bajas voluntarias e involuntarias o mantenerlas en conjunto. También si se evaluará la posibilidad de hacer algún tipo de clustering para mejorar el poder de predicción. De ser puesto en producción, el modelo deberá estar siendo continuamente actualizado debido a los cambios en los datos para los siguientes meses. Esto también involucra una ventana de tiempo entre los datos de entrenamiento y la validación, que será la implementación real del modelo. Habría que encontrar cuál es la mejor ventana de tiempo para ver cuánto tiempo tener para tomar acción antes de la baja, así como también cuántos meses de historia se deberían tener en cuenta.

Los datos

Los datos del banco contienen todos los movimientos del cliente. Por cuestiones de confidencialidad, los datos serán anonimizados para no dañar la privacidad de la empresa que brindó los mismos.

Entre ellos se encuentran 5 fuentes principales.

La fuente OS, con 294 variables, que indica la información de los seguros del último mes, en particular de las pólizas que el cliente tiene con el banco.

La fuente OSA_3M, con 22 variables, que indica las altas de seguros del último mes y agregaciones de los últimos 3.

La fuente OSB_12M, con 150 variables, que indica las bajas de seguros del último mes y agregaciones de los últimos 3, 6 y 12 meses.

La fuente MAPA_3M, con 9 variables, que indica el mapa del último mes y agregaciones de los últimos 3.

La fuente CL_PL_3M, con 1862 variables, que tiene la información de clientes del último mes y agregaciones generales de los últimos 3 y algunas puntuales más extensas.

En principio lo que se haría serían agregaciones por cliente y tipo de seguro. Cruzando los datos de cada cliente por número de identificación, se podrán hacer ingeniería de variables para poder implementar los distintos modelos y luego elegir el mejor.