INSTITUTO TECNOLÓGICO DE BUENOS AIRES – ITBA

ESCUELA DE (INGENIERÍA Y TECNOLOGÍA – INGENIERÍA Y GESTIÓN - POSTGRADO)

MODELO PREDICTIVO DE CANCELACION DE DONACIONES EN ORGANIZACIONES SIN FINES DE LUCRO

El caso de DonarOnline.org

AUTOR/ES: Aran, María Inés (Leg. N° 104124)

DOCENTE/S TITULAR/ES O TUTOR/ES: Vaisman, Alejandro

TRABAJO FINAL PRESENTADO PARA LA OBTENCIÓN DEL TÍTULO DE ESPECIALISTA EN CIENCIA DE DATOS

BUENOS AIRES

SEGUNDO CUATRIMESTRE, 2019

1. Introducción

La ciencia de datos y el big data se han desarrollado principalmente en el sector privado, donde se han implementado técnicas estadísticas y de machine learning que han permitido el aumento de los beneficios económicos y una mayor eficiencia en el uso de recursos en las empresas.

El sector social en Argentina, conformado por organizaciones sin fines de lucro, puede beneficiarse con la aplicación de las nuevas tecnologías vinculadas a los datos de manera tal que también les permitan mejorar sus procesos internos y generar mayor impacto.

En este sentido resulta relevante explorar la aplicación de modelos de machine learning a problemáticas del sector social.

Donaronline.org es una sociedad de la sociedad civil (OSC) de Argentina que presta servicio a otras OSC ayudándolas a gestionar el cobro de donaciones en línea con tarjeta de crédito o débito.

La recopilación sistemática de datos referidos a las donaciones recibidas y de las OSC beneficiarias permiten entrenar un modelo de machine learning para predecir la cancelación de donaciones.

Un modelo que prediga la cancelación de donaciones permite a las organizaciones sin fines de lucro identificar a los donantes a los que dirigir campañas de retención de donaciones. Por otro lado, da visibilidad respecto de los aspectos que una organización sin fines de lucro debe tener en cuenta para disminuir el riesgo de perder donaciones.

2. Estado de la cuestión

El desafío de retener clientes (en el caso del sector privado) o donantes (en el caso del sector social) ha permitido el desarrollo de distintas herramientas estadísticas con el objetivo de generar alertas tempranas de una probable baja de la compra o donación.

En el caso del sector social, retener donantes es relevante para mantener fondos estables para poder financiar los proyectos que llevan a cabo (Barber (4); Bennet (5)).

La incorporación de técnicas de machine learning para resolver estos problemas y específicamente en el sector social es mas incipiente. En este sentido Althoff (3) estudian los factores que inciden en que los donantes se mantengan en la plataforma de financiación colectiva. Detectan que la distancia geográfica entre el donante y la organización a la que ayudan es significativa para definir si volverán a donar. También la cantidad de datos que el donante brinda respecto de si mismo esta relacionada con la voluntad de volver a donar y la comunicación del impacto de la donación es crucial para aumentar la fidelidad. Para predecir la probabilidad de que vuelva a donar, los autores modelan una regresión logística con un resultado de 0.74 de AUC de la curva ROC.

Schetgen (9) analiza la importancia del uso de las redes sociales por parte de las Organizaciones sin fines de lucro como herramienta para comprender mejor

las características de sus donantes. En este trabajo se generan tres modelos con el objetivo de predecir la probabilidad de que una persona que es fan en Facebook se convierta en donante de la organización. El algoritmo que mejor resultado arroja es random forest con un AUC de 0.66.

En cuanto a los motivos que llevan a los donantes a mantener su donación son diversos. A. ([2], [1]); Sargeant A. ([8]) identifican factores importantes para la retención de donantes incluyendo la construcción de relaciones, la comunicación del impacto, la confianza, el compromiso, la satisfacción y la participación del donante en las actividades que lleva a cabo la organización.

Lu ([7]) analizan la aplicación de análisis de supervivencia para estimar el valor de vida de los clientes de una empresa de telecomunicaciones. Utilizan el análisis de supervivencia para modelar una regresión logística, validan los resultados eligiendo un determinado momento en el futuro y ordenando las probabilidades de supervivencia predichas por el modelo para cada uno de los clientes. Luego, agrupan los casos en deciles y encuentran que los tres primeros deciles contienen el 74 % de los clientes que sobreviven (siguen siendo clientes) al tercer mes.

En el trabajo de África Periáñez and C. ([10]) se aplica un ensamble de arboles de análisis de supervivencia para predecir el tiempo que pasara hasta que un jugador de un juego online abandone la plataforma. Se comparan la utilización de técnicas tradicionales de análisis de supervivencia como la regresión Cox con un ensamble de arboles de análisis de supervivencia obteniendo un error de predicción (Integrated Brier Score de 0.158) para el segundo método y un error de 0.169 (IBS) utilizando una regresión Cox.

3. Planteamiento del problema

3.1. Operaciones de Donaronline

Donaronline.org es una organización de la sociedad civil de Argentina que presta servicios a otras organizaciones sin fines de lucro (OSC u ONG) en América Latina, ayudándolas a gestionar las donaciones con tarjeta de crédito y débito que reciben, simplificando al máximo posible el proceso, desde la intención de la donación hasta que la misma llega a la organización.

En el caso de Argentina las donaciones se gestionan a través de dos canales: Mercadopago y CentralPos.

Donaronline recauda un aporte mensual del 2,7 % por cada transacción efectiva que la ONG recibe y que fue realizada a través de DonarOnline. Si la ONG recibe una donación mensual menor o igual a $ 2000 o U$S 112 los servicios de la plataforma Donaronline son gratuitos.

3.2. Proceso de donación

El donante prospecto ingresa a la página web de la ONG o a un enlace web compartido en redes sociales y es redirigido a una página web de Donaronline que contiene el formulario. En el mismo, completa datos personales tales como

nombre, apellido, DNI, domicilio, e-mail, entre otros, además de los datos del método de pago: tarjeta de crédito o débito.

Una vez completado el formulario, se re-direcciona al donante a una pagina web en la que se agradece la donación y se invita al donante a compartir un mensaje en redes sociales, invitando a otras personas a donar a la ONG.

Adicionalmente, el donante recibe un correo electrónico con los detalles de la donación y un mensaje personalizado por cada ONG en el que detalla las actividades que realiza la organización y un agradecimiento.

3.3. Problema

Donaronline cuenta con datos de las donaciones que reciben las ONGs que realizan campañas a través de su pagina web. Actualmente, se almacenan los datos de todo el proceso por el que atraviesa el donante: desde que el potencial donante ingresa al sitio web de las ONGs hasta que se vuelve efectiva la transacción monetaria.

Estos datos se registran con el fin de mantener una trazabilidad del proceso de donación y de facilitar la infraestructura y la gestión del cobro a las ONG.

Existe un potencial de mejora y enriquecimiento del servicio ofrecido por Donaronline, basado en el desarrollo de productos de análisis de esos datos y de identificación temprana de posibles cancelaciones de donaciones. El objetivo es que la predicción pueda ser utilizada por las ONGs para campañas de fidelización y comunicación temprana que permitan disminuir la cantidad de donaciones perdidas.

3.4. Modelo de datos

Donaronline utiliza un modelo de datos relacional.

La tabla de hechos (donations) almacena todas las intenciones de donación que se registran a través del formulario en internet.

La tabla donations se relaciona con la tabla donors en las que se almacenan todos los datos personales que las personas ingresaron en el formulario de donación. Además, se registra si el donante compartió en redes sociales (Facebook o Twitter) un mensaje que invita a otras personas a donar a la causa a la que acaba de contribuir (Esta acciones quedan registradas en la tabla shares).

Las tablas payment_methods y payment_transactions contienen los pagos recibidos y rechazados de cada donación.

La tabla campaigns contiene los datos de la campaña a la que el donante esta contribuyendo. Para cada campaña, en la tabla analytics, se registran todos los ingresos que hubo al formulario (ya sea que se hayan convertido en donación o solo visita) y la pagina desde donde fueron direccionados al formulario: Facebook, búsqueda de Google, Pagina web de DonarOnline o pagina web de la ONG.

Las campañas son llevadas a cabo por las organizaciones cuyos datos se almacenan en la tabla organizations. La categoría a la que la ONG pertenece (medio ambiente, pobreza, niñez, entre otras) se almacena en las tablas tags y

tagging. Por otro lado, los usuarios con perfiles con permiso para administrar las campañas de las ONGs se almacenan en la tablas users y profiles. La tabla versions contiene información respecto de los cambios que los usuarios hicieron tanto en los perfiles de la ONG a la que representan como en la actualización de los datos de sus donantes.

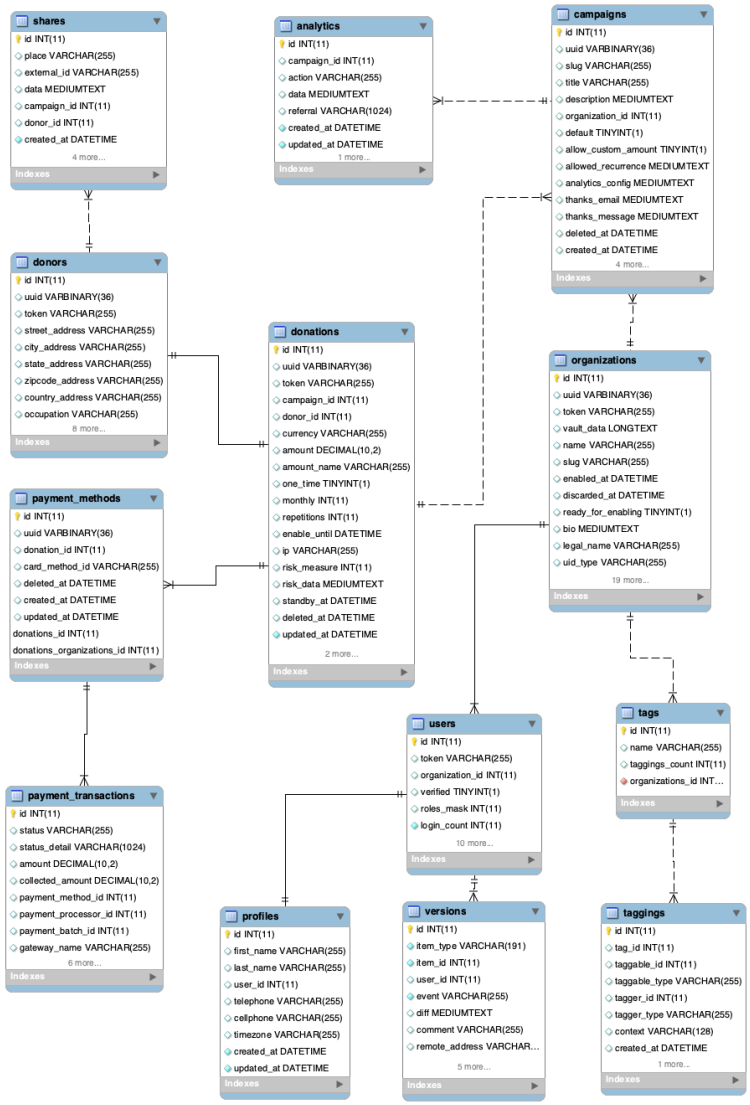


Figura 1: Modelo de datos

4. Solución

El objetivo es determinar si un modelo de machine learning permite predecir la cancelación de donaciones mensuales en las ONG.

Se construye un conjunto de datos basado en ventanas temporales que abarca desde 30 de Noviembre de 2012 hasta 23 de Julio de 2018 para poder entrenar y evaluar distintos modelos.

Se generan modelos con tres algoritmos y distintas combinaciones de hiperparametros.

Las métricas empleadas para evaluar los resultados son el area debajo de la curva ROC y el valor-F.

4.1. Conjunto de datos de entrenamiento

El conjunto de datos de entrenamiento contiene los datos a nivel donación de la tabla donations y agrega información de sus tablas relacionadas.

La variable a predecir es la baja de la donación, para eso se consideran solo las donaciones con recurrencia mensual. Esta variable toma valor 1 si la donación fue dada de baja y 0 en caso contrario.

El conjunto de entrenamiento contiene 233 variables contenidas en las categorías que se describen a continuación. El anexo contiene más detalle de las variables generadas.

Variables relativas a la donación Los datos que surgen del formulario de donación tales como la organización, campaña a la que esta contribuyendo y el monto mensual a contribuir. Además se generan los campos de día, horario de donación y una variable binaria para identificar si la fecha de registro de donación fue en la primera o segunda quincena del mes.

Variables relativas a los donantes Los registros de la tabla donors se utilizan para generar variables binarias que indiquen si el donante ha completado sus datos de teléfono y celular, si la ONG a la que contribuye ha actualizado sus datos personales y la cantidad de días transcurridos desde la última actualización. Además se contabilizan la cantidad de campos completados por el donante.

Variables relativas a la campaña Estas variables consideran la duración (en días) de la campaña, una variable binaria que indica si se envía un mensaje de agradecimiento a los donantes por su registro en el formulario de donación y las variables relativas a los datos que surgen de la tabla analytics.

Se generan variables relativas a la cantidad de tráfico web del formulario de donación de cada campaña proveniente de cada uno de los canales (Facebook, Google, DonarOnline, página web de la ONG). Además se construyen variables que reflejen la variación en la cantidad de visitas recibidas al formulario en el tiempo.

Variables relativas a la ONG Los datos relativos a la ONG y su actividad tales como la categoría a la que pertenecen y si tienen pagina de Facebook, Twitter y/o Linkedin.

Variables relativas a las transacciones monetarias Estas variables registran la cantidad y montos recaudados por cada donación, el medio de pago y la cantidad de rechazos o transacciones que no pudieron ser cobradas. Se generan variables adicionales que contabilizan la cantidad de transacciones rechazadas sobre el total de transacciones, la cantidad de medios de pagos distintos utilizados, entre otros.

Variables relativas al comportamiento de pago Estas variables calculan el promedio de número de día de la semana en la que se registran las transacciones aprobadas y rechazadas.

4.2. Validación cruzada utilizando ventanas temporales

Los problemas de predicción requieren tomar en consideración el tiempo. Para generar el conjunto de datos para entrenar el modelo se tomaron ventanas temporales de forma tal que simularan la generación de datos del modelo en producción.

Las ventanas temporales toman como punto de referencia una determinada fecha. A partir de ella, se calculan las variables explicativas hasta el día anterior a la fecha de referencia ("ventana de observación"). El resultado de la variable a predecir (cancela o no la donación) se computa a partir de la fecha de referencia ("ventana de resultados").

A modo de ejemplo, tomando de referencia el 31 de Enero de 2018, la "ventana de observación" abarca las donaciones que iniciaron entre Noviembre 2017 y 31 de Enero 2018 para las cuales se calculan todas las variables explicativas al 30 de enero de 2018.

La "ventana de resultados" que se inicia a partir del 01 de Febrero de 2018, abarca el periodo Febrero 2018 a Marzo 2018 y contiene la variable que indica si la donación fue dada de baja en ese periodo.

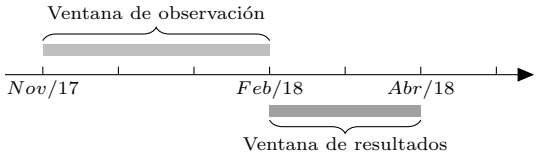


Figura 2: Construcción de ventanas temporales de datos de entrenamiento

Se repite este procedimiento corriendo la fecha de referencia cada sesenta días. De esta manera quedan formadas 32 ventanas temporales. Los modelos se entrenan utilizando una ventana temporal y se evalúan en la siguiente. Este

procedimiento se repite 32 veces y se deja de lado la ultima ventana que se utilizará como conjunto de datos para validación.

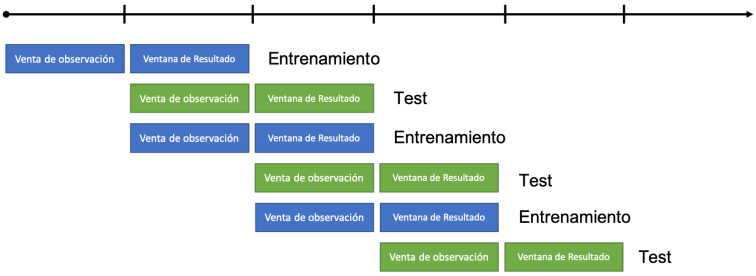


Figura 3: Validación cruzada temporal

4.3. Herramientas utilizadas

La base de datos relacional se almacena en PostgreSQL 10. La generación del conjunto de datos de entrenamiento y validación utilizando la validacion cruzada con ventanas temporal se realiza a través de consultas de SQL ejecutadas a traves de Python 3.6.

Para la geolocalización de los datos se generó un script en Python 3.6 utilizando el paquete geopy.

Los modelos se construyen en R versión 3.4.1 y su evaluación y selección se hace en Python 3.6.

4.4. Modelos

Los conjuntos de datos de entrenamiento se utilizan para entrenar modelos que pueda identificar patrones que determinan la probabilidad de que una donación sea dada de baja en el futuro.

Se entrenan con los métodos de regresión logística y 'random forest' y 'extreme randomized trees' con distintas combinaciones de hiperparametros.

Regresión logística El método de regresión logística predice la probabilidad de que una observación(en esta caso una donación) sea cancelada utilizando variables independientes continuas. Se utiliza la forma funcional logística que para un determinado conjunto de variables independientes predice valores entre 0 y 1.

$$log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X \tag{1}$$

El modelo utiliza los datos de entrenamiento para estimar los coeficientes β. Para eso se aplica el método de minimos cuadrados ordinarios que consiste en seleccionar aquellos valores de β que minimicen el error entre el valor predicho y el real.

Random forest El algoritmo de método supervisado Random Forest (Bosques aleatorios) para modelos de regresión evalúa los datos de entrenamiento, crea arboles de decisión y los ensambla para otorgar a cada donación la probabilidad de que sea dada de baja.

A partir de muestras aleatorias con reposición de las observaciones del conjunto de entrenamiento se generan arboles de decisión. La estructura de cada árbol esta conformada por nodos para los que se selecciona una variable y se prueban distintos puntos de corte seleccionando aquel que separa mejor las donaciones que fueron dadas de baja de las que siguen activas.

Extra trees: Extremely randomized trees Es una metodología similar a la de bosques aleatorios pero que difieren en dos aspectos.

El primero es que no utilizan las muestras aleatorias con reposición para generar los arboles de decisión.

El segundo es que es que para seleccionar una valor de corte para dividir cada nodo, primero se seleccionan una cantidad mínima de valores de forma aleatoria y luego entre ellos se escoge el mejor valor.

Selección de hiperparámetros Los hiperparámetros se refieren a configuraciones manuales externas al modelo y que no se estiman utilizando los datos. Cada algoritmo contiene hiperparámetros que pueden configurarse para mejorar los resultados de la predicción. Ya que la mejor configuración no es conocida a priori, se prueban distintas combinaciones y se seleccionan aquellos hiperparámetros que mejor resultado arrojen.

En el caso de la regresión logística solo se entrenó con hiperparámetros por defecto de la librería 'caret' de R. Los algoritmos de 'random forest' y 'extremely randomized trees' comparten algunos hiperparámetros tales como 'mtry'(cantidad de variables disponibles para dividir en cada nodo del árbol), 'ntree' (cantidad de arboles a construir) y 'nodesize'(cantidad mínima de observaciones en los nodos terminales). Los arboles extremadamente aleatorios también utilizan el hiperparámetro 'numRandomCuts' que refiere a la cantidad de valores aleatorios a utilizar para cortar un nodo.

Los hiperparámetros que se probaron fueron los siguientes:

- mtry: 186,15,77
- ntree: 500,1000,1500
- nodesize: 2,4,6
- numRandomCuts: 1,2,6

4.5. Entrenamiento de los modelos

Los datos de entrenamiento de las distintas ventanas temporales se modelan utilizando los tres algoritmos y las distintas combinaciones de hiperparámetros.

Se calcula el área debajo de la curva ROC como medida de evaluación del rendimiento.

El proceso funciona de la siguiente manera: Se toma una ventana temporal para entrenar y se generan distintos modelos utilizando los distintos algoritmos y sus configuraciones de hiperparámetros. Luego, cada uno de los modelos es aplicado a la siguiente ventana temporal de test y se calcula el área debajo de la curva ROC (AUC) como medida de evaluación del rendimiento del modelo.

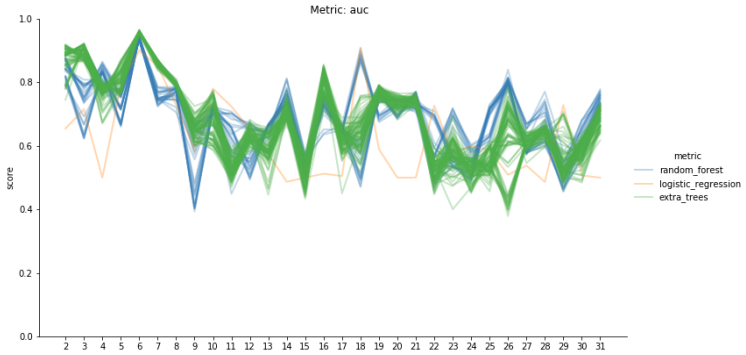


Figura 4: Resultado de modelos para cada ventana temporal de evaluación

En el gráfico de la figura 4 se observa el resultado de cada modelo(combinación de algoritmo e hiperparámetros) para cada ventana temporal, el color identifica el algoritmo con el que se entreno.

La mayoria de los modelos generados mantienen un rendimiento entre 0.5 y 1 de AUC en cualquiera de las ventanas temporales, lo que implica un resultado mejor que la aleatoriedad.

5. Validación

5.1. Selección del mejor modelo

En todos los modelos generados se calculan el área debajo de la curva ROC y la métrica valor-F para ser utilizadas como medidas de rendimiento.

Métricas Las métricas que se utilizan para evaluar los distintos modelos son el área debajo de la curva ROC y el valor-F.

Todos los modelos generan predicciones entre 0 y 1 para determinar la probabilidad de que una donación sea dada de baja. Los puntos de corte se utilizan como valores divisorios para determinar que las probabilidades por debajo de ese valor sean clasificadas como 'alta probabilidad de cancelación' y lo contrario en el caso de que el valor predicho sea mayor al del punto de corte.

La curva ROC es la representación de la especificidad y la sensibilidad para distintos puntos de corte.

Dado que los casos de cancelación de donaciones representan solo el 3 % del total, es de interés calcular la métrica Valor-F que considera tanto la precisión como la exhaustividad.

$$Valor - F = 2\frac{Precisión * Exhaustividad}{Precisión + Exhaustividad} \tag{2}$$

$$Precisión = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ positivos} \tag{3}$$

$$Exhaustividad = 2\frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ negativos} \tag{4}$$

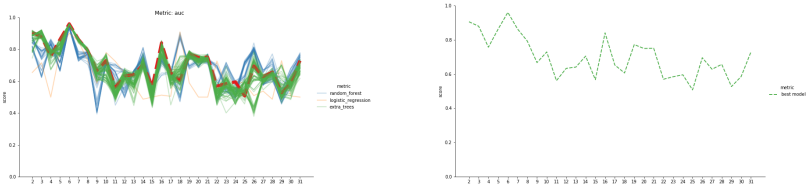
A fin de poder calcular la métrica Valor-F, para cada modelo en cada ventana temporal se seleccionó el punto de corte óptimo (método optimalCutoff) de la librería InformationValue de R.

El mejor modelo De todos los modelos generados se escoge el mejor, de acuerdo a los siguientes criterios:

1. Selección de aquellos modelos cuyo AUC sea mayor a 0.5 en todas las ventanas temporales de evaluación.
2. Selección de los diez mejores modelos ordenados de forma descendente de acuerdo al promedio de AUC contabilizando todos las ventanas temporales de evaluación.
3. Selección del que tenga mayor promedio de valor-F contabilizando todas las ventanas temporales de evaluación.

El modelo seleccionado es el construido con el algoritmo 'extremely randomized trees' y la siguiente combinación de hiperparámetros:

- mtry: 15
- ntree: 1500
- nodesize: 2
- numRandomCuts: 1



(a) Resultado de modelos para cada ventana temporal

(b) Resultado de mejor modelo para cada ventana temporal

Figura 5: Comparación del mejor modelo encontrando respecto de todos los modelos entrenados

El AUC promedio para todas las ventanas temporales es de 0.7 y el valor-F promedio es de 0.35.

5.2. Datos de validación

La evaluación de la capacidad predictiva del modelo construido se determina aplicando el modelo seleccionado a un nuevo conjunto de datos que en este caso corresponde a la ventana temporal de evaluación mas reciente(ventana 32) que se dejo de lado en el entrenamiento y la selección de los modelos.

Se analiza la cantidad de aciertos y errores de predicción del modelo para esta nueva ventana temporal.

El conjunto de datos contiene 2.6 % de cancelaciones respecto del total de donaciones del periodo. Mantiene proporciones similares a las ventanas temporales de entrenamiento.

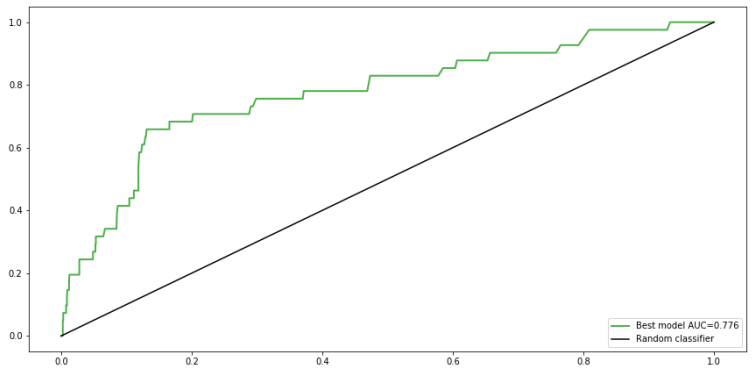


Figura 6: Área debajo de la curva ROC para el conjunto de datos de validación

El área debajo de la curva ROC para el conjunto de datos de validación es de 0.77.

Variables relevantes Las variables más relevantes para predecir si una donación sera cancelada en el futuro están relacionadas con las variables relativas a campañas llevadas a cabo por las OSCs y el análisis de su tráfico web.

La cantidad promedio visitas que el formulario de donación recibió en los últimos 12, 6 y 3 meses es relevante para predecir que la donación sea dada de baja (Analytics_Camp_Views_Avg_Year, Analytics_Camp_Views_Avg_Semester, Analytics_Camp_Views_Avg_Quarter).

La cantidad de visitas recibidas en el ultimo periodo también es una variable con capacidad predictiva (Analytics_Camp_Views_T). En lo que se refiere a canales de comunicación por los que los visitantes del formulario ingresaron, las visitas recibidas desde Facebook son las más relevantes para identificar la probabilidad de cancelación de la donación.

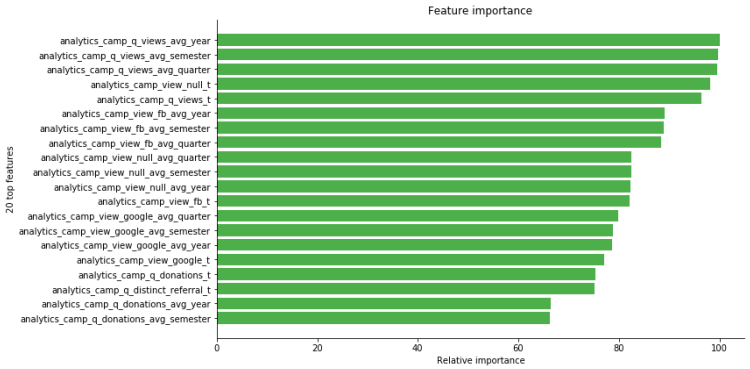


Figura 7: Variables más importantes para predecir la baja de donantes

6. Conclusiones y futuros trabajos

Los datos de los donantes, su comportamiento de pago de donaciones y la actividad de las Organizaciones de la Sociedad Civil en redes sociales e internet permiten predecir la probabilidad de que una donación sea dada de baja.

El problema de predicción de cancelación de donaciones es un problema de datos desbalanceados, donde la clase a predecir representa el 3 % del total de los datos.

El modelo desarrollado con el algoritmo 'extremly randomized trees' arroja resultados de un área debajo de la curva de 0.77 para el conjunto de datos de validación. Estos resultados permiten considerar el modelo desarrollado como una buena herramienta para la predicción de cancelación de donaciones.

Las variables generadas para enriquecer el modelo relacionadas con la actividad de las organizaciones en redes sociales resultaron importantes para la discriminación entre cancelaciones y no cancelaciones. Estos resultados son similares a los obtenidos por (9). Por otro lado, las variables referidas a la cantidad de datos completados en el formulario y la distancia geográfica entre la organización y el donante no resultaron relevantes, estos resultados son contrarios a los hallazgos de (3).

Las organizaciones de la sociedad civil se pueden ver beneficiadas con la incorporación de técnicas de machine learning que les permitan volverse más eficientes. El uso de este modelo para predecir la baja de donantes les permitiría realizar campañas para retener a aquellos donantes con altas probabilidades de darse de cancelación.

En el futuro, es de interés indagar en otras técnicas estadísticas y de machine learning que permitan además de predecir la baja de la donación, estimar el momento el que esta ocurrirá. Para eso existen extensiones de la técnica random forest para incorporar el análisis de supervivencia.

A. Apendice

A.1. Conjunto de datos de entrenamiento

El conjunto de datos de entrenamiento contiene 233 variables. A continuación se describe con detalle las variables generadas en la ingeniería de atributos.

Variables relativas a los donantes: Completitud del formulario La cantidad de datos personales que el donante brinda a la OSC a la cual dona le permite a esta última estar en contacto con sus donantes y saber más respecto de las características de las personas. Se genera un atributo que mide la cantidad de datos completos por el donante al llenar el formulario.

El objetivo es identificar si los donantes que mas datos completan son los mas comprometidos y en consecuencia su probabilidad de darse de baja es menor que aquellos que completan menos datos.

$$Completitud_i = \frac{QC_i}{QT} \tag{5}$$

Donde QC_i indica la cantidad de datos completos por el donante i y QT la cantidad total de campos del formulario de donación.

Variables relativas a los donantes: Georeferenciación Los donantes pueden completar en el formulario de donación el dato de domicilio, ciudad y provincia. En los casos en los que esta información haya sido provista, se puede identificar la latitud y longitud del donante y de las ONG utilizando el servicio de API Open Source de (Foundation).

Con los datos georefenciados se calcula la distancia en linea recta entre la dirección declarada por el donante y la de la ONG.

Esta variable permite estudiar la hipótesis de que a mayor distancia física mayor es la probabilidad de baja de donación.

Variables relativas a campañas: Análisis de tráfico web Se generan variables a nivel campaña que reflejan la tendencia de cantidad de visitas y donaciones recibidas a través de cada uno de los canales por los que los visitantes y donantes fueron dirigidos al formulario.

Estas variables se utilizan como aproximación de la actividad de las organizaciones sin fines de lucro en las redes sociales y paginas web.

$$Analytics\_Camp\_Q\_Distinct\_Referral = Q_ct \tag{6}$$

Donde Q_ct indica la cantidad de canales distintos (Facebook, Sitio web de la OSC, sitio web de DonarOnline, búsqueda en Google u otro) por los que la campaña c recibió visitas al formulario de donaciones en el momento t.

$$Analytics\_Camp\_Q\_Distinct\_Referral\_Dif\_T\_T1 = \frac{Q_ct}{Q_ct_{-1}} \tag{7}$$

Donde $Q_c t_{-1}$ indica la cantidad de canales distintos (Facebook, Sitio web de la OSC, sitio web de DonarOnline, búsqueda en Google u otro) por los que la campaña c recibió visitas al formulario de donaciones en el momento $t-1$.

$$Analytics\_Camp\_Ratio\_Donation = \frac{Q_c d}{Q_c v} \tag{8}$$

Donde $Q_c d$ indica la cantidad de veces que se completo el formulario y $Q_c v$ la cantidad de visitas que el formulario recibió.

$$Analytics\_Camp\_View\_Avg\_F = \frac{\sum_{i=1}^{N} Q_i}{f} \tag{9}$$

Donde Q_i indica la cantidad de veces que se visito el formulario de la OSC en el mes i. F indica la frecuencia, en el caso del año(Year) F toma valor 12, 6 en el caso de semestre y 3 en el caso de trimestre.

Variables relativas al comportamiento de pago: Promedio de día de pago Se toman todas las transacciones efectivas recibidas para cada donación y se selecciona el día en que fueron pagadas. Se calcula el día de pago promedio.

$$Avg\_pay\_day_i = \frac{\sum_{i=1}^{N} d_i}{N} \tag{10}$$

Donde d_i indica el día en el que fue pagada la donación y N la cantidad de veces que se recibieron transacciones efectivas para la donación i.

Bibliografía

[1] A., S. (2008). Donor retention: What do we know and what can we do about it? *A Report for the Association of Fundraising Professionals*.

[2] A., S. (2011). Relationship fundraising: How to keep donors loyal. *Nonprofit Management and Leadership*.

[3] Althoff, Tim y Leskovec, J. (2015). Donor retention in online crowdfunding communities: A case study of donorschoose.org. *Proceedings of the 24th International Conference on World Wide Web*.

[4] Barber, Putnam y Levis, B. (2013). Donor retention matters. *Center on nonprofits and Philanthropy - Urban Institute*.

[5] Bennet, R. (2006). Predicting the lifetime durations of donors to charities. *Journal of Nonprofit Public Sector Marketing*.

[Foundation] Foundation, O. Data extracted from openstreetmap after september 2012 is licensed on terms of the open database license, .ºdbl"1.0, previously it was licensed cc-by-sa 2.0.

[7] Lu, Junxiang y Park, O. (2019). Modeling customer lifetime value using survival analysis - an application in the telecommunications industry.

[8] Sargeant A., Ford J.B., W. D. (2006). Perceptual determinants of nonprofit giving behavior. *Journal of Business Research*.

[9] Schetgen, L. (2017-2018). Predicting donation behavior: Acquisition modeling in the nonprofit sector based on facebook data. Master's thesis, Universitet Geint, Belgica.

[10] África Periáñez, Saas A., G. A. and C., M. (2016). Churn prediction in mobile social games: Towards a complete assessment using survival ensembles. *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*.