**Especificación del modelo de ML**

1. Motivación del proyecto de ML

¿Cuál es el problema específico de ML que se quiere resolver?: Predicción de futuros clientes que estén propensos a cancelar una póliza de vehículo

¿Cómo este problema se articula con el problema de analítica formulado?: En la actualidad la empresa que está prestando los datos para el estudio no cuenta con una herramienta que permita identificar los factores que llevan a un retiro, por eso se necesita un algoritmo predictivo que permita cumplir con este objetivo y así implementar estrategias de retención de manera preventiva

1. Definición del problema

¿Cuál es la salida especifica que se va a predecir?: Una probabilidad que indique si la persona se retirará o no, o un estimado de en cuánto tiempo (meses, años) se retirará el cliente de acuerdo a ciertas características

¿Cuáles son los datos de entrada requeridos para el proceso de analítica?: Base de datos con más de 4000 registros y doce variables en formato Excel, incluyendo una categoría con la etiqueta de si son retirados o activos

Para el dataset a utilizar en ML describa:

1. Número de registros de entrenamiento: En total se presentan 4753 registros, pero de ellos se tomará el 70% para la parte de entrenamiento, que correspondería aproximadamente a 3327 registros
2. Existen datos no-balanceados, existen sesgos en los datos: No, si bien la etiqueta de los datos que indica un retiro no se encuentra dividida en un 50-50, si hay una proporción lo suficientemente equitativa como para no implementar un algoritmo de oversampling o una metodología similar para el balanceo de los datos
3. ¿Qué tan frecuentemente se actualizan los datos?: La última actualización se presentó en Mayo, no obstante una vez construida e implementada la herramienta la idea es recibir actualización de la base cada 3 meses

¿Cuáles son los factores más relevantes que permitirían predecir la salida?: Tener una base lo suficientemente limpia para evitar complicaciones en la construcción del algoritmo, conocer claramente el contexto de la problemática y así delimitar los objetivos y posibles caminos a recorrer. Contar con información suficiente para el análisis también es importante, y en lo posible sin datos faltantes o inconsistencias significativas

1. Medidas de desempeño

¿Se tiene un marco de trabajo experimental de línea base?

Si:

¿Existe documentación?

¿Cómo se construyó?

¿Cuál es su desempeño y cómo se mide?

No:

¿Cuál es el marco experimental propuesto (que experimentos de predicción se requieren)? Se desea realizar un modelo de sobrevivencia o de regresión de Cox para predecir el tiempo en que una persona puede dar de baja su póliza, dependiendo del desempeño del modelo se plantea una segunda opción en caso de requerirla que es un modelo de regresión logística o random forest para clasificar a las personas como posibles a retirarse o no en base al histórico de datos y a las etiquetas que se tienen

¿Cómo se va a medir el desempeño del modelo de ML? La evaluación del desempeño de un modelo de supervivencia para predecir datos futuros es la consideración más importante, es por esto que las métricas a emplear en este caso son las proporciones de riesgo o *Hazard ratio*, las diferencias significativas en las curvas de supervivencia de Kaplan-Meier de los dos grupos., también se utilizarán el coeficiente de determinación R2, la correlación de rango de Somers y el área bajo la curva ROC (AUC)

¿Cuál son las implicaciones de tener FP y FN (errores)? Se tiene conocimiento de que el modelo no será 100% exacto por lo que si se presentan errores habrá clientes que probablemente no se retiren en el tiempo estimado, presentándose realmente antes o después de lo que arroja la predicción, no obstante, es una consecuencia natural del modelo

¿Cuál es una buena solución?: Implementar métodos adecuados en la partición, evaluar cuáles son las variables que realmente aportan a la predicción y en general probar todos los modelos posibles para encontrar el más adecuado y que tenga una mejor precisión y performance

¿Cuál es la literatura relacionada al problema (ej. papers y/o proyectos)?: En cuánto a técnicas clásicas de ML para el manejo de esta problemática en la industria aseguradora se encuentran los siguientes:

* Clara-Cecilie Gunther, Ingunn Fride Tvete, Kjersti Aas, Geir Inge Sandnes, and

rnulf Borgan. Modelling and predicting customer churn from an insurance company. Scandinavian Actuarial Journal, 2014(1):58-71, February 2014.

* W. Verbeke, D. Martens, C. Mues, and B. Baesens. Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Systems with Applications, 38(3):2354-2364, 2011. ISSN 09574174. doi: 10.1016/j.eswa.2010.08.023. URL: <http://dx.doi.org/10.1016/j.eswa.2010.08.023>.

Implementando exclusivamente modelos de análisis de supervivencia se encuentran los siguientes:

* P. L. Brockett, L. L. Golden, M. Guillen, J. P. Nielsen, J. Parner, and A. M. Perez-Marin. Survival analysis of a household portfolio of insurance policies: how much time do you have to stop total customer defection? Journal of Risk and Insurance, 75(3):713{737, 2008.
* M. Henao, D. Restrepo and H. Laniado. Customer churn prediction in insurance industries: a multiproduct aproach. Master Tesis. Universidad EAFIT. March 4, 2020
* D. Van den Poel and B. Lariviere. Customer attrition analysis for financial services using proportional hazard models. European journal of operational research, 157(1):196-217, 2004.

1. Partición de los datos

¿Qué consideraciones se deben tener en cuenta para la partición de los datos (sesgos, imbalance de clases, múltiples etiquetas)? Para este caso se tiene una variable binaria para las etiquetas, que es la que indica si la persona está retirada o activa, por lo que puede implicar un proceso más sencillo para la predicción. Por otro lado se considera que no se presenta un imbalance significativo en esta variable respuesta ya que la proporción de retirados es de un poco más del 38% por lo cual no se empleara una técnica de oversampling o algo similar para tratar este problema ya que por ahora no se considera necesario

¿Qué estrategia de partición se va a utilizar (random, LOOCV, K-folding)?: Se utilizará K-fold cross validation

1. Línea de tiempo

¿Cuándo es el deadline? No hay una fecha específica en que deba estar desarrollada la herramienta, pero entre más pronto esté más rápido se puede empezar a adaptar e implementar

¿Cuándo se deben los primeros resultados? Probablemente a mitad de agosto

¿Cuándo se debe tener la solución propuesta? Al menos en los próximos tres meses

1. Contactos

¿Quién es responsable del proyecto? El área de data & analytics de la compañía Aon Risk Services

¿Quién es el responsable del acceso a los datos (en training y test)? Los datos fueron enviados por parte de la compañía interesada en el estudio por lo que cualquier inquietud y demás será directamente con ellos

¿Quién puede ayudar a entender el proceso, los resultados (experto de dominio)? El área de data & analytics de la compañía Aon Risk Services y el área Affinity para todo el aspecto comercial de cara al cliente

1. Recursos para entrenamiento

¿Cuáles son los recursos computacionales (software/hardware) disponibles para entrenamiento? Cloudera para emplear R y Power bi para la presentación de resultados

¿Cuál es el presupuesto (tiempo/dinero) disponible para entrenamiento? En tiempo será alrededor de 3 meses

1. Despliegue

¿Cuáles son las restricciones tecnológicas para el despliegue del modelo (software particular, maquinas particulares)? Hasta el momento no encontramos ninguna restricción en este aspecto para el desarrollo de la metodología

¿Cuál va a ser el uso del modelo (consultas diarias, mensuales, etc)? Consultas trimestrales

¿Cuál va a ser el mecanismo de acceso al modelo? Mediante un tablero de Power bi que presente las métricas y visualizaciones necesarias para conocer los resultados del modelo

¿Cómo se va a integrar el modelo con otros componentes del negocio (personas otros sistemas)? El tablero se encontrará en un micrositio de la compañía para así posteriormente el asesor y experto comercial encargado del cliente revise los resultados y diseñe las estrategias correspondientes