

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Modelo de Supervivencia para predecir la permanencia de los colaboradores en la empresa Claro Colombia

---

INFORME DE PASANTÍA PARA OPTAR POR EL TÍTULO DE MATEMÁTICA
PROGRAMA ACADÉMICO DE MATEMÁTICAS

Angie Lorena Moreno Matiz
Profesional Designado: Lina María Garzón Méndez
Docente Director: Luis Alejandro Masmela Caita

Bogotá DC
Julio de 2023

Resumen

El presente informe, titulado "Modelo de Supervivencia para predecir la permanencia de los colaboradores en la empresa Claro Colombia" tiene como propósito exponer el trabajo desarrollado durante mi pasantía en Claro Colombia cuyo objetivo fue implementar un modelo que estime el tiempo que va a permanecer cada colaborador en la compañía. Esta estimación servirá como base para la formulación de planes de acción orientados a la gestión de la rotación del talento humano.

El proceso se llevó a cabo mediante un análisis de supervivencia, que empleó los métodos Kaplan-Meier y Regresión de Cox. Estos métodos proporcionaron resultados altamente específicos en relación con la duración proyectada de la permanencia de cada individuo en la organización. Además, se identificaron las variables y factores de mayor influencia en estas estimaciones. A partir de este análisis, fue posible identificar de manera precisa el grupo de individuos con mayor riesgo de abandonar la empresa en un futuro cercano. Estos hallazgos serán fundamentales para la implementación de planes estratégicos de fidelización y retención, con el propósito de fortalecer la relación con estos colaboradores y aumentar su permanencia en la organización.

Índice

1. Introducción **3**

2. Propuesta de Pasantía **3**

2.1. Objetivos 3

2.1.1. Objetivo General 3

2.1.2. Objetivos Específicos 4

2.2. Cronograma 4

3. Preliminares ***5***

4. Metodología **7**

4.1. Primer acercamiento 7

4.2. Procesamiento de datos 8

4.3. Estructuración del modelo y resultados 9

4.4. Evaluación 12

5. Discusión **13**

5.1. Socialización 15

6. Conclusiones **15**

7. Agradecimientos **15**

8. Anexos **17**

8.1. Librerías utilizadas en Python 17

1. Introducción

Este informe presenta la pasantía desarrollada en Claro Colombia, una destacada empresa de telecomunicaciones que opera en el país desde 1994 bajo la razón social COMCEL, ofreciendo una amplia gama de servicios innovadores en el ámbito de las comunicaciones. Como parte del grupo América Móvil, una de las empresas de telecomunicaciones más grandes a nivel global, Claro Colombia se ha posicionado como líder en el mercado local y ha contribuido significativamente al avance de la conectividad en todo el país[1].

En un contexto de continuo avance y adaptación a las metodologías ágiles, hacia el año 2020, Claro Colombia implementó la Gerencia de Transformación y People Analytics de la Dirección de Gestión Humana. La iniciativa de People Analytics consiste en la aplicación de técnicas de análisis de datos y estadísticas para comprender mejor el comportamiento, las tendencias y las características de los empleados en una organización para modernizar la gestión de los recursos humanos dentro de la misma. Esto implica la recopilación y análisis de información relacionada a los empleados con el objetivo de tomar decisiones basadas en los datos [9].

En esta iniciativa se establece el objetivo de la pasantía, el cual es ajustar un modelo para estimar el tiempo de permanencia de cada colaborador en la compañía con la finalidad de generar un insumo para la fidelización de los colaboradores y evitar altos niveles de rotación.

Los colaboradores de Claro Colombia se agrupan principalmente en tres categorías según el tipo de contrato: directos, temporales y aprendices. El enfoque del modelo predictivo se centró exclusivamente en los colaboradores de tipo directo.

Para el desarrollo de la pasantía se utilizó la aplicación Microsoft Excel [1] para el almacenamiento y la gestión de los datos y, el lenguaje de programación Python [6] para el desarrollo del modelo.

2. Propuesta de Pasantía

2.1. Objetivos

2.1.1. Objetivo General

Ajustar un modelo de supervivencia a un conjunto de datos de los colaboradores de Claro Colombia, para predecir el tiempo que permanecerán vinculados a la compañía.

[1] Más información en: `https://www.elempleo.com/sitios-empresariales/colombia/claro-colombia/quienes_somos.asp`

2.1.2. Objetivos Específicos

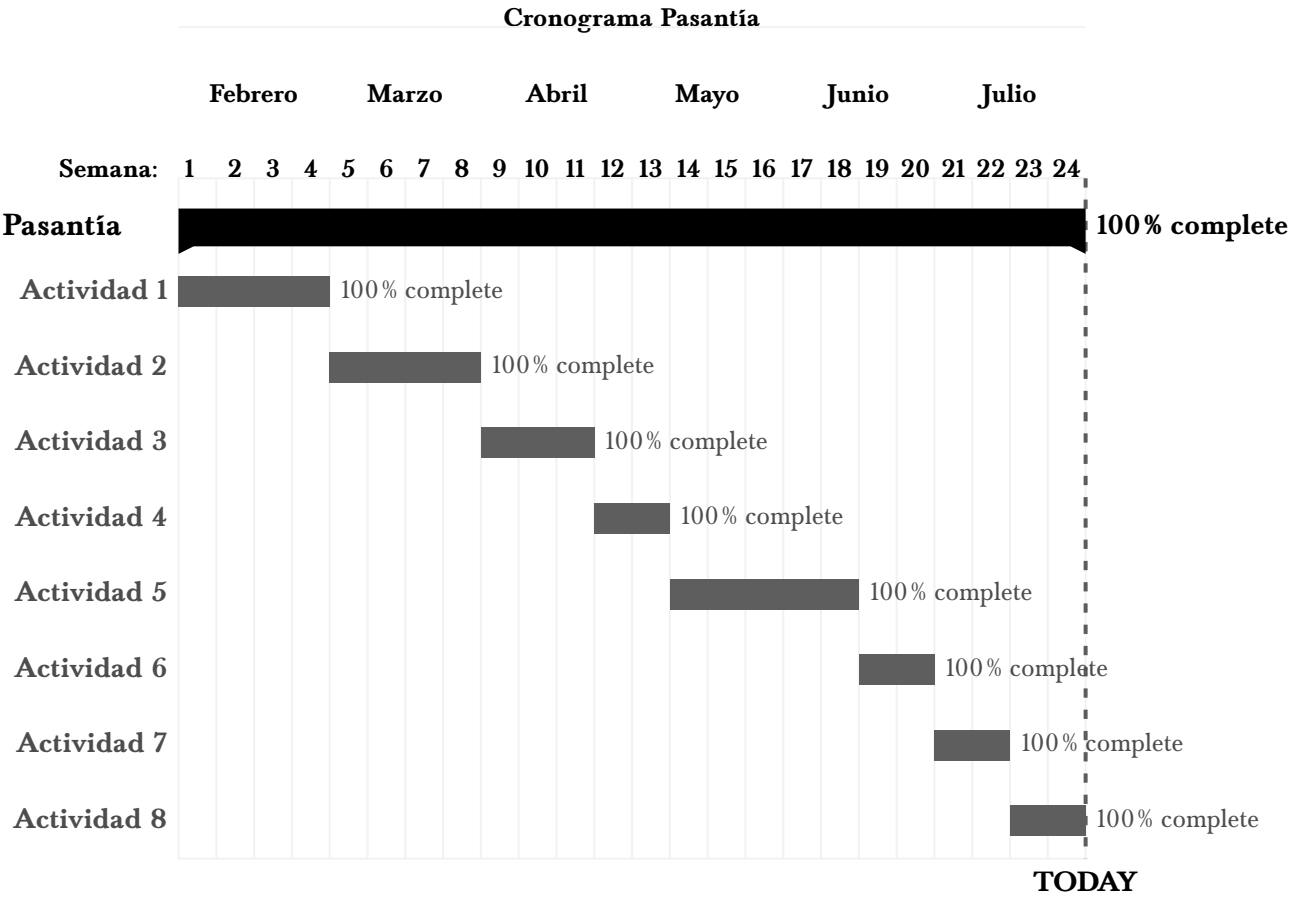
- Realizar una revisión bibliográfica sobre modelos de supervivencia y su aplicación en el ámbito empresarial.
- Identificar y analizar las variables relevantes que pueden influir en la permanencia de los colaboradores en la empresa.
- Llevar a cabo el procesamiento de los datos.
- Realizar un análisis descriptivo y exploratorio de los datos para identificar posibles tendencias y patrones.
- Seleccionar y evaluar la calidad del modelo final y, determinar si es apropiado para su implementación en la empresa Claro Colombia.
- Generar recomendaciones prácticas para la empresa en función de los resultados obtenidos.

2.2. Cronograma

La pasantía tuvo una duración de 6 meses, en la cual se plantearon las siguientes actividades para el cumplimiento de los objetivos:

1. Estudio Modelo Predictivo de Fuga de Claro Colombia.
2. Revisión bibliográfica Modelos de Supervivencia.
3. Estudio variables.
4. Procesamiento de datos.
5. Ajuste Modelos de Supervivencia.
6. Evaluación modelo (estudio de capacidad predictiva).
7. Análisis de resultados.
8. Socialización del modelo en Claro Colombia.

Estas actividades se distribuyeron de la siguiente manera:



3. Preliminares

La necesidad de saber el tiempo que va a permanecer cada colaborador en la compañía conlleva a realizar un análisis de supervivencia el cual es una técnica estadística utilizada para analizar el tiempo transcurrido hasta que se produce un suceso específico que puede ocurrir a lo largo del tiempo como la muerte, la recuperación de una enfermedad, la finalización de un proyecto o, en este caso, la deserción. La aplicación del análisis de supervivencia en la gestión de recursos humanos no es una novedad, ya que uno de sus propósitos más frecuentes consiste en anticipar la rotación. La predicción de la rotación se ha convertido en un reto fundamental en las organizaciones, ya que sus implicaciones abarcan no solo pérdidas económicas, sino también un detrimento en el capital humano disponible [21].

A diferencia de otros tipos de análisis estadísticos que se basan en observaciones independientes, el análisis de supervivencia considera el tiempo de ocurrencia del suceso y tiene en cuenta el hecho de que algunos individuos pueden no experimentar el suceso antes del final del estudio [10].

Dentro del análisis de supervivencia existen distintos métodos [12, p. 7] :

- Métodos Análisis de Supervivencia
  - No Paramétricos
    - Kaplan-Meier
    - Nelson-Aalen
    - Life-Table
  - Semi-paramétricos
    - Regresión de Cox
  - Paramétricos
    - Regresión Lineal
    - Accelerated Failure Time
  - Machine Learning
    - Árboles de Supervivencia
    - Métodos Bayesianos
    - Redes Neuronales
    - Support Vector Machine
    - Ensemble
    - Machine Learning Avanzado

Los métodos no paramétricos son más eficaces cuando no se conoce o no es evidente la distribución del tiempo de supervivencia, en los métodos semi-paramétricos no es relevante esta información y, en los métodos paramétricos el tiempo de supervivencia tiene alguna distribución en particular [12, p. 8].

Los métodos usados con más frecuencia son Kaplan-Meier y Regresión de Cox. La **estimación de Kaplan-Meier** sirve para darse una idea preliminar del comportamiento de las variables individualmente. Por otra parte, **la Regresión de Cox**, también llamada Modelo Proporcional de Cox o Modelo de Cox, es el modelo más utilizado en el análisis de supervivencia puesto que incorpora covariables, es decir, es multivariado y estima el impacto de cada variable en la supervivencia [14], [17].

Para cada uno de estos métodos existen conceptos básicos a tener en cuenta en el conjunto de datos [11, p. 3], estos son:

- **Evento**: En el análisis de supervivencia, el evento es el acontecimiento de interés en el estudio, para este caso, la deserción de los colaboradores.
- **Tiempo**: Es el tiempo transcurrido desde el inicio de un periodo de observación hasta un acontecimiento o el final del estudio.

- **Censura**: La censura se produce cuando no se conoce exactamente el tiempo de supervivencia particular. El individuo está censurado si no se observa ni se sabe nada sobre él después del momento de la censura. Un sujeto censurado puede o no tener un suceso tras el final del tiempo de observación.

En el lenguaje de programación Python [6] se utiliza la librería `scikit-survival` [15] para el análisis de supervivencia.

4. Metodología

4.1. Primer acercamiento

En la iniciativa People Analytics ya se manejaba un modelo predictivo de fuga/churn de talento en el lenguaje de programación Python [6], el cual indica la probabilidad que tiene cada persona de irse de la compañía y los clasifica en muy baja, baja, media, alta o muy alta probabilidad de fuga. Este cálculo lo realiza haciendo uso de distintas técnicas de análisis predictivo tales como redes neuronales artificiales, las cuales hacen parte del machine learning[2] y realizan un proceso llamado " aprendizaje profundo" que utiliza nodos o neuronas de manera semejante a como funciona el cerebro humano y de tal modo se va perfeccionando continuamente [18]; regresión logística, que es una técnica estadística utilizada a menudo para el modelado y la analítica predictiva, su uso más frecuente es predecir la probabilidad de que ocurra un evento o de que se tome una decisión [8]; entre otras.

Cada uno de estos métodos se ejecuta y generan una asignación de pesos a cada variable incluida en el modelo para finalmente determinar la probabilidad que tiene cada persona de irse de la compañía. Al realizar este proceso, se comparan todas las técnicas y se identifica cuál tiene mayor porcentaje de predicción.

Estudiar este modelo fue necesario para crear un primer acercamiento a las bases de datos y a la información que se tenía disponible. Los resultados que habían producto de la última ejecución arrojaban que aproximadamente 2.500 personas tenían alta o muy alta probabilidad de fuga (superior al 75 %), un número que está muy lejos del índice de rotación tanto mensual como anual de la compañía. A partir de esto surgió un gran cuestionamiento: Si una persona tiene alta o muy alta probabilidad de fuga, ¿cuándo se materializará? Esto teniendo en cuenta que aunque una persona desee abandonar la compañía, pueden existir diversos factores que se lo impidan. Debido a esto, se propuso el objetivo de la pasantía.

[2] Rama de la inteligencia artificial (IA) y la ciencia de la computación que se centra en el uso de datos y algoritmos para imitar la forma en que aprenden los humanos, mejorando gradualmente su precisión [7].

4.2. Procesamiento de datos

Teniendo en cuenta que ya se ha iniciado un estudio sobre la deserción de los colaboradores en la compañía y el modelo propuesto tiene un objetivo similar, se tomó la decisión de usar las mismas bases de datos con el fin de tener homogeneidad en la información para más adelante poder comparar y hacer un cruce de ambos modelos para generar mayor valor al estudio.

Estas bases de datos incluyen:

- Información personal y demográfica: Nombre completo, cédula, fecha de nacimiento, edad, antigüedad (en años), género, deserción (hace referencia a activo o retirado), dirección de residencia, número de hijos, pasatiempos, tipo de vivienda, estado civil, estrato socioeconómico, lugar de nacimiento, entre otros.
- Estudios: Nivel de estudio culminado, en curso, cursos obligatorios de la compañía, cursos voluntarios y horas de formación.
- Información laboral: Cargo, dirección de comité, región, nombre del jefe, sindicatos, proyectos a los que pertenece, salario, beneficios, entre otros.
- Indicadores: Resultados de encuesta NSU, clima, evaluación 360.

Para un total de 222 columnas y 14306 filas inicialmente. Una vez actualizada toda la información se procedió con el procesamiento y limpieza de datos[3] en Python.

En primera instancia se verificó el porcentaje de campos vacíos por columna y no se tuvieron en cuenta 9 variables que carecían de más del 50 % de los datos.

Luego, se evidenció que algunas columnas contenían campos con la misma información escrita de distintas maneras. Se procedió a unificar la escritura de estas categorías, por lo cual fue necesario corregirlos de tal manera que quedara un solo registro. Hecho esto, se verificó el porcentaje de datos en 0 por columna y se identificó en qué casos esto representaba falta de información, con lo cual se prescindió de 13 variables.

El paso a seguir fue codificar las variables categóricas a variables binarias o "dummies" para que el modelo capture correctamente el efecto de cada categoría en la supervivencia a lo largo del tiempo.

Finalmente, el estado de censura, es decir, la variable "Deserción" se establece como booleana. De esta manera se da por concluido el procesamiento y limpieza de datos. Obteniendo así, un conjunto de más de 200 variables únicamente de carácter numérico o binario (a excepción de la variable deserción) que serán en las que se basará el modelo para aprender comportamientos y a partir de ello generar la predicción.

[3] La limpieza de datos es un conjunto de pasos para mejorar la calidad de la información. Se trata de corregir y eliminar datos incorrectos en una base de datos, lo cual implica encontrar y cambiar datos incompletos, incorrectos, dañados o que no son útiles [4].

4.3. Estructuración del modelo y resultados

Una vez lista la información y culminado el estudio bibliográfico de análisis de supervivencia, se toma la decisión de iniciar dicho análisis con la estimación de Kaplan-Meier para ver cómo están relacionadas una o dos variables en específico con la deserción de los colaboradores y luego, en base a ello, continuar con el Modelo Proporcional de Cox [14], [17].

Se procede con la instalación de paquetes y librerías necesarias en Python[4].

Hecho esto, se inicia el análisis de supervivencia con el **Estimador de Kaplan-Meier**. Este método se usa para estimar cuántas personas sobreviven a lo largo del tiempo.

La representación gráfica de esta función es conocida como la curva de Kaplan-Meier. Esta curva muestra la probabilidad de que ocurra un evento en un cierto periodo, como la deserción. Si el grupo estudiado es lo suficientemente grande, la curva debería acercarse a cómo realmente sobreviven las personas en la población. Para calcular la probabilidad de sobrevivir en un momento específico, se divide el número de personas que siguen vivas en ese momento entre el número total de personas en riesgo. Las personas cuya información está incompleta no se consideran en el total.

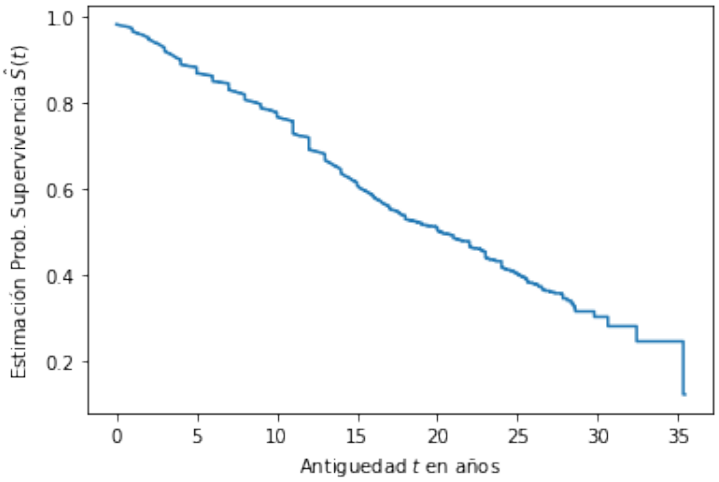


Figura 1: Curva Kaplan-Meier según antigüedad

La curva que se presenta en la Figura 1 se obtuvo a partir del cruce de las variables antigüedad (tiempo) y deserción (evento), en la cual se observa cómo cambian las probabilidades de supervivencia a lo largo del horizonte temporal: mientras más antigüedad, se reducen las probabilidades de supervivencia.

También se estimó la mediana del tiempo de supervivencia y los intervalos de confianza del 95 %. La mediana obtenida fue de 20 años, lo que indica que el 50 % de la muestra permanece en la compañía

[4] Véase section 8

al menos 20 años y el 50 % se va en ese tiempo. Los límites del intervalo de confianza del 95 % tanto inferior como superior fueron 19.3 y 21 años, respectivamente.

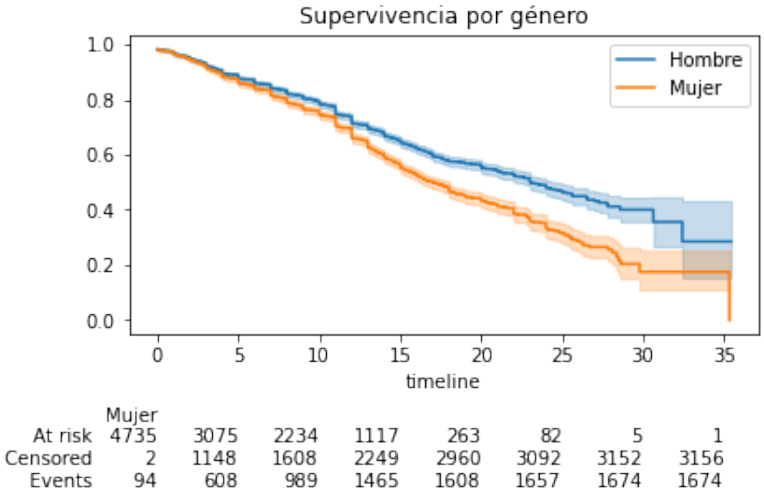


Figura 2: Curva Kaplan-Meier según género

La curva de Kaplan-Meier también se puede generar discriminada a partir de variables categóricas. Por ejemplo, la Figura 2 muestra la curva de Kaplan-Meier para cada una de las categorías de la variable género. En ella que se observa que el género masculino tiene una mayor probabilidad de supervivencia.

Una vez visto cómo funciona el estimador de Kaplan-Meier, se procedió con el **Modelo de Cox**. Este modelo, como se mencionaba anteriormente, considera múltiples variables al mismo tiempo y analiza cómo se relacionan con la distribución de supervivencia.

Para un ajuste exitoso del modelo fue necesario eliminar la multicolinealidad[5], puesto que su existencia afecta de manera negativa al modelo. Este proceso se llevó a cabo utilizando el **Factor de Inflación de la Varianza** o **VIF** por sus siglas en inglés, el cual permite cuantificar la intensidad de la multicolinealidad. Este valor siempre es positivo y crece a medida que aumenta la multicolinealidad entre las características.

En este modelo se consideró un VIF elevado por encima de 10, es decir, al ejecutarse la función en Python, se calcula el VIF a todo el conjunto de datos, en cuanto encuentra que una columna tiene VIF superior a 10, la elimina y vuelve a hacer el cálculo con el nuevo conjunto de datos. El proceso se detiene cuando no se encuentra un VIF superior a 10.

[5]La multicolinealidad es una situación que se da cuando entre dos o más variables de un modelo de regresión existe una relación muy fuerte, lo cual reduce la rigurosidad de la estimación [20].

A continuación, teniendo en cuenta el alto volumen de características con las que se contaban, se vió la necesidad de añadir una penalización para ajustar el modelo. El objetivo de añadir esta regularización es poner una restricción a la magnitud de los coeficientes del modelo.

Para llevar a cabo esta penalización, se estudiaron los métodos **Ridge**, **Lasso** y **Elastic Net** [2] para luego definir cuál es el que mejor se adaptaba al modelo [16].

- **Ridge**: Esta regresión es conocida por su capacidad para manejar la multicolinealidad, controlar el sobreajuste y la magnitud de los coeficientes, esto lo hace disminuyéndolos a valores cercanos a 0 sin llegar a ser iguales a 0. Además, funciona mejor cuando la mayoría de los atributos son relevantes [2].

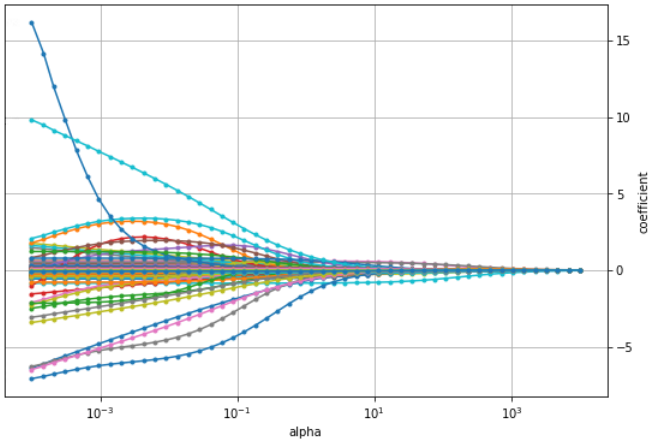


Figura 3: Penalización de Ridge

La Figura 3 representa el efecto de esta penalización en las variables utilizadas en el modelo.

- **Lasso**: Esta regresión no penaliza de la misma manera a los coeficientes demasiado grandes. Mientras que en la regresión Ridge los coeficientes tienden hacia 0, aquí los coeficientes pueden volverse 0, es decir, se eliminan automáticamente las variables que no agregan poder predictivo al modelo y de esta manera se puede decir que se realiza una selección automática de características. Es importante tener en cuenta que Lasso funciona mejor cuando no existe alta multicolinealidad [2].

  La Figura 4 representa el efecto de esta penalización en las variables utilizadas en el modelo.

- **Elastic Net**: Esta regresión es una combinación entre Ridge y Lasso y es efectiva cuando hay un gran número de atributos. Sin embargo, esto no significa que sea siempre la mejor opción [2].

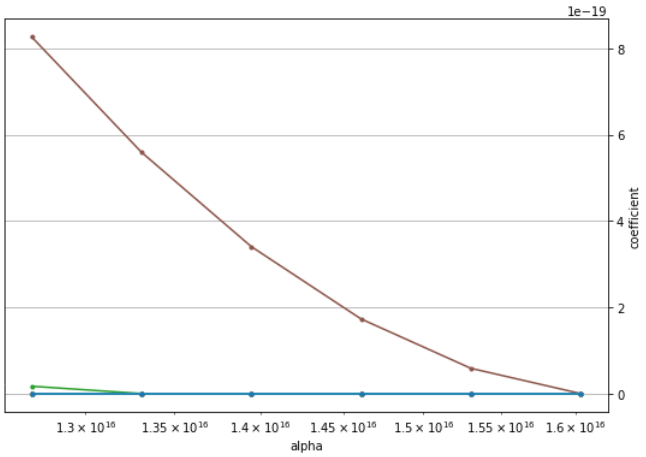


Figura 4: Penalización de Lasso

Para el caso de este modelo, el comportamiento de Elastic Net fue bastante similar al de Lasso (Figura 4).

Por tal motivo, es necesario validar cuál coeficiente de penalización es el mejor haciendo un cruce entre estos métodos. El resultado obtenido, es decir, el valor que el programa nos indica que es el mejor parámetro penalizador fue 0.0032.

En este punto ya se pudo ajustar el Modelo de Cox. Una vez ejecutado el modelo, el paso a seguir era extraer el tiempo que se estimaba que iba a permanecer cada persona en la compañía. Para esto se usó la función `cph.predict_expectation` [5], la cual genera el valor en específico del tiempo que se predice que va a permanecer cada colaborador en la compañía.

Esta estimación se dió en años y se añadió como nueva columna a la base de datos para finalmente exportarla nuevamente a Excel.

Adicionalmente, del modelo en Python se pueden extraer los coeficientes de cada una de las variables, en donde un coeficiente positivo indica que esa variable se asocia a un aumento de la probabilidad de deserción y, por el contrario, un coeficiente negativo se asocia a la disminución del riesgo de deserción. Esto quiere decir que se pueden conocer cuales características tuvieron mayor efecto en la predicción del tiempo de supervivencia de cada persona.

4.4. Evaluación

Dado que se está trabajando con un conjunto de datos censurados, el modelo no se puede evaluar con un método regular como el error cuadrático [12, p. 21], por el contrario, se debe hacer con técnicas específicas para el análisis de supervivencia [13].

Para este modelo, la evaluación de la capacidad predictiva se realizó en términos de la concordancia entre los tiempos de supervivencia que ya se tenían (antigüedad) y los predichos por el mismo. Esto se hace mediante el **análisis de Concordancia de Harrell** también conocido como **C-index**, cuyo valor varía entre 0 y 1 donde 1 indica una concordancia perfecta.

Al hacer esta evaluación, se obtuvo un C-index de 0.9901, el cual indica una buena capacidad predictiva del modelo realizado durante la pasantía. Dados estos resultados, se determinó que el Modelo de Cox sería un buen complemento para el Modelo de Fuga.

5. Discusión

En el archivo final que se exportó a Excel se realizó la comparación del tiempo estimado según el Modelo de Cox y la antigüedad de cada persona con el fin de visualizar con facilidad el "tiempo restante" de cada uno en la compañía. Este tiempo restante se clasificó en: corto plazo (menos de 1 año), mediano plazo (entre 1 y 2.5 años) y, largo plazo (más de 2.5 años). Esta distribución se hizo con la intención de generar la relación planteada con el modelo de fuga ya existente en la gerencia como se observa en la Figura 5.

La manera en la que se interpreta la Figura 5 es según la escala de color: un tono de rojo más fuerte indica que a esa categoría o grupo de personas se les predice un riesgo de fuga más alto en un plazo de tiempo más cercano, es decir, son las que mayor alerta generan. A medida que va disminuyendo la tonalidad de rojo, disminuye también la alerta generada por cada grupo de personas. Esto hasta llegar a una tonalidad amarillo claro, la cual indica un riesgo de fuga más bajo en un tiempo más alejado, por lo cual causa una alerta más leve sin que esta deje de existir.

De manera más puntual, las personas que tienen muy alta probabilidad de fuga y se estima que van a desertar en menos de un año son las que mayor alerta generan (rojo intenso) y son a las que hay que designarles un plan de acción para que no abandonen la compañía. Por otra parte, las personas que tienen muy baja probabilidad de fuga con un tiempo estimado de deserción mayor a 2.5 años son las que menor alerta generan (amarillo claro) y, por el momento, no representan un riesgo.

Además, como se mencionaba anteriormente, del modelo se pueden conocer las variables que más influyen en la deserción de los colaboradores y en las cuales se deben focalizar los planes de acción para la fidelización de los mismos. En este caso, el modelo arrojó que estas variables son:

- Antigüedad menor a 4 años: esta causa se vinculó directamente a la rotación temprana[6] de colaboradores.

[6]La rotación temprana es el porcentaje de empleados que dejan un trabajo dentro de su primer año u otro período definido por la organización [3].

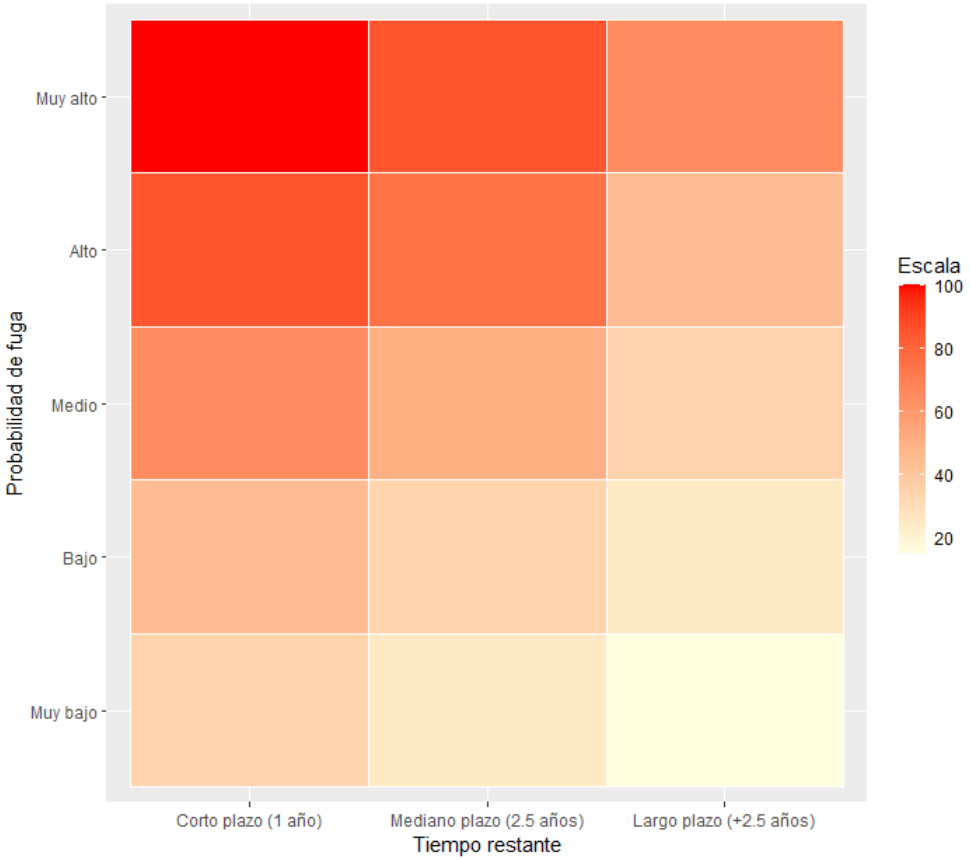


Figura 5: Relación Modelo de Fuga y Modelo de Cox

- Salario: aquí se enlazan varios factores como la equidad salarial[7] tanto a nivel de la organización como del mercado, el tipo de salario (fijo o variable), entre otros.

Finalmente, del cruce ilustrado en la Figura 5 de los dos modelos, se pudo caracterizar a las personas con mayor riesgo, es decir, las variables que más se repiten en este grupo de personas. De tal manera, se obtuvo que las personas que mayor alerta generan son de género masculino, entre 30 y 40 años de edad, solteros, sin hijos, con altos niveles de educación (posgrados), mayor formación adicional como cursos obligatorios y libres, ingresados recientemente a la compañía y con inconformidades con respecto al salario.

[7]La equidad salarial implica considerar diversos factores, como habilidades, experiencia, responsabilidades, entre otros, con el fin de determinar una remuneración justa y proporcional para cada persona [19].

5.1. Socialización

De esta manera, se socializó el Modelo de Cox a la Gerencia Transformación y People Analytics obteniendo comentarios positivos no solo por la efectividad del modelo sino también porque es una propuesta innovadora en el joven proyecto de People Analytics tanto en Claro Colombia como a nivel latinoamericano. Luego de la aprobación de la gerencia, se realizó la socialización con la Dirección de Gestión Humana, siendo invitadas las gerencias a las cuales les podría ser de utilidad el modelo, como por ejemplo la Gerencia Gestión del Talento, entre otras. Allí el modelo tuvo un amplio recibimiento y, gracias a él, se definió la creación de un nuevo comité llamado Comité de Fidelización el cual tiene por objetivo evitar llegar a la retención del colaborador, sino por el contrario, fidelizarlo antes de que desee abandonar la compañía generando planes de acción focalizados principalmente en las variables que el Modelo de Cox establece como las que tienen mayor influencia en la deserción de los colaboradores. Este comité se propuso realizarlo mensualmente, lo que implica a su vez la actualización y ejecución mensual del modelo.

6. Conclusiones

- Se consultó la teoría que soportan diversos modelos de supervivencia en la búsqueda de un modelo que lograra predecir el tiempo de supervivencia de cada colaborador.
- Se identificó desde el conjunto de variables que se usó en el modelo de fuga, aquellas variables significativas para el modelo de supervivencia.
- Se realizó el proceso de limpieza de la base de datos lo que permitió trabajar con una base de datos final que contenía 14306 individuos y 200 variables.
- Fueron utilizados los métodos Kaplan-Meier y Regresión de Cox para realizar el análisis de la información.
- Se evaluó el modelo final mediante el Índice de Concordancia de Harrel el cual arrojó un C-index de 0.99, lo que sugiere que es altamente confiable.
- Se extrajeron las variables que más influyen en la deserción de los colaboradores para focalizar los planes de acción necesarios para prevenir la rotación.

7. Agradecimientos

En primer lugar, quiero expresar mi profundo agradecimiento a mi familia por su apoyo constante a lo largo de mi trayectoria. Mi gratitud también se extiende a los docentes, cuyo compromiso de

compartir sus conocimientos en cada clase superó cada uno de los obstáculos. Finalmente, gracias a Claro Colombia, por brindarme la oportunidad de formar parte del equipo de practicantes Generación XXI. Estos seis meses representaron un período de crecimiento significativo y amplio aprendizaje, además de ser una primera incursión en el mundo laboral.

Referencias

[1] Microsoft 2023®. Microsoft excel. https://www.microsoft.com/es-co/microsoft-365/excel.

[2] Bautista Luis Alberto. Regresión penalizada: Ridge, lasso y elastic net. 8 2023.

[3] Rh Campus. Rotación temprana. https://rhcampus.net/blog/rotacion-temprana/#:~:text=La%20tasa%20de%20rotaci%C3%B3n%20temprana,le%20pide%20que%20se%20vaya).

[4] DataScientest. Datacleaning limpieza de datos: definición, técnicas, importancia en data science. https://datascientest.com/es/datacleaning-limpieza-de-datos-definicion-tecnicas-importancia-en-data-science, abril 2022.

[5] Cam Davidson-Pilon. Coxphfitter. https://lifelines.readthedocs.io/en/latest/fitters/regression/CoxPHFitter.html#coxphfitter.

[6] Python Software Foundation®. Python. https://www.python.org/.

[7] IBM. What is machine learning? https://www.ibm.com/topics/machine-learning.

[8] IBM. ¿qué es la regresión logística? https://www.ibm.com/mx-es/topics/logistic-regression.

[9] Spiceworks Inc®. What is people analytics? definition, process, trends, and tools. https://www.spiceworks.com/hr/hr-analytics/articles/what-is-people-analytics/, marzo 2021.

[10] Investopedia. Survival analysis. https://www.investopedia.com/terms/s/survival-analysis.asp, agosto 2021.

[11] Avishek Nag. *Survival Analysis with Python.* Auerbach Publications, 1 edition, 2021.

[12] Chandan K. Reddy Ping Wang, Yan Li. Machine learning for survival analysis: A survey. *ACM Computing Surveys*, 51(6):110, 2 2019.

[13] Sebastian Pölsterl and contributors. Evaluating survival models. httpEvaluatingSurvivalModelss://scikit-survival.readthedocs.io/en/stable/user_guide/evaluating-survival-models.html.

[14] Sebastian Pölsterl and contributors. Introduction to survival analysis with scikit-survival. https://scikit-survival.readthedocs.io/en/stable/user_guide/00-introduction.html.

[15] Sebastian Pölsterl and contributors. scikit-survival. https://scikit-survival.readthedocs.io/en/stable/.

[16] Sebastian Pölsterl and contributors. scikit-survival. https://scikit-survival.readthedocs.io/en/stable/user_guide/coxnet.html.

[17] Rahul Raoniar. Survival analysis in python (km estimate, cox-ph and aft model). https://medium.com/the-researchers-guide/survival-analysis-in-python-km-estimate-cox-ph-and-aft-model-5533843c5d5d, octubre 2021.

[18] Amazon Web Services®. ¿qué es una red neuronal? https://aws.amazon.com/es/what-is/neural-network/.

[19] Great Place To Work®. Equidad salarial vs igualdad salarial. https://www.greatplacetowork.com.co/es/recursos/blog/equidad-salarial-vs-igualdad-salarial, junio 2023.

[20] Probabilidad y Estadística. Multicolinealidad. https://www.probabilidadyestadistica.net/multicolinealidad/#google_vignette.

[21] Qianwen Zhu, Jiaxing Shang, Xinjun Cai, Linli Jiang, Feiyi Liu, and Baohua Qiang. Coxrf: Employee turnover prediction based on survival analysis. pages 1123–1130, 08 2019.

8. Anexos

8.1. Librerías utilizadas en Python

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import xlrd
import sqldf
```

```
from scipy.stats import norm
import sys
from collections import Counter
from statsmodels.stats.outliers_influence import variance_inflation_factor


from datetime import datetime, date, time, timedelta
import fuzzywuzzy
from fuzzywuzzy import fuzz
from fuzzywuzzy import process


from lifelines import CoxPHFitter
from lifelines import KaplanMeierFitter
from lifelines.utils import median_survival_times
from lifelines.utils import survival_table_from_events
from sksurv.nonparametric import kaplan_meier_estimator
from sksurv.linear_model import CoxPHSurvivalAnalysis, CoxnetSurvivalAnalysis
from sksurv.preprocessing import OneHotEncoder


from sklearn import set_config
from sklearn.model_selection import GridSearchCV, KFold
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler


from lifelines.utils import concordance_index
```