INFORME CASO DE ESTUDIO ANALÍTICA FINANCIERA

ELABORADO POR:

ANDRÉS FELIPE PENNA HERNÁNDEZ

JHONATAN VALENCIA OCAMPO

STEVEN ZAPATA ZULETA

PROFESOR:

ANDRES MAURICIO GOMEZ ARDILA

ANALÍTICA

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

a. Diseño de solución propuesto.

Problema de negocio

La tarificación precisa es un factor crítico para el éxito de cualquier compañía de seguros. Una tarificación incorrecta puede tener graves consecuencias, incluyendo la pérdida de clientes y la disminución de ganancias. Si las tarifas son demasiado altas, los asegurados pueden buscar alternativas más económicas o incluso decidir no adquirir un seguro oncológico en absoluto, lo que resultaría en una pérdida de oportunidades y una disminución en la base de clientes. Por otro lado, si las tarifas son demasiado bajas, la compañía puede enfrentar pérdidas financieras significativas si los costos médicos relacionados con el cáncer exceden las primas recolectadas.

El modelo analítico propuesto permitirá a UdeA Insurance calcular de manera precisa y personalizada los costos médicos para cada asegurado, teniendo en cuenta variables relevantes como la edad, el historial médico, entre otros. Esto garantizará una tarificación justa y adecuada, ajustada a las necesidades y características individuales de cada asegurado.

Además, la interpretabilidad y transparencia del modelo son aspectos fundamentales para fortalecer la confianza de los clientes. Al poder explicar claramente a los asegurados cómo se determina su tarifa, UdeA Insurance demostrará un enfoque justo y basado en criterios objetivos, lo que aumentará la confianza en el proceso de tarificación. Esta transparencia fortalecerá la relación con los clientes, fomentando la fidelidad y el reconocimiento de la compañía como líder en el mercado de seguros oncológicos.

Problema analítico

Se establecen los siguientes apartados de analítica, estos son:

1. Definir las variables relevantes que permitan diseñar un sistema de predicción de tarifas al seguro oncológico que se base en los datos obtenidos por parte del usuario.
2. Establecer un modelo de regresión que se ajuste a las variables elegidas, esto se valida de acuerdo con su desempeño.
3. Realizar un modelo de tarife de forma precisa el seguro oncológico (calculando una prima justa), basado en perfilamiento individual y sus riesgos.

Marco

La tarificación precisa de un seguro oncológico es de suma importancia para UdeA Insurance, ya que afecta directamente tanto a los asegurados como a la compañía. Una tarificación

incorrecta puede tener consecuencias negativas significativas. Si la tarifa es demasiado alta, los asegurados pueden optar por no adquirir el seguro o buscar alternativas más asequibles, lo que resulta en una pérdida de oportunidades comerciales y clientes potenciales. Por otro lado, si la tarifa es demasiado baja, la compañía puede enfrentar pérdidas financieras considerables si los costos médicos relacionados con el tratamiento del cáncer superan las primas recaudadas.

Al desarrollar un modelo analítico para la tarificación del seguro oncológico, UdeA Insurance busca evitar estos escenarios indeseables. La precisión en la determinación de los costos médicos individuales y el perfilamiento de los asegurados permite establecer tarifas acordes a su nivel de riesgo. Al contar con un modelo interpretable y transparente, la compañía puede explicar a los clientes cómo se determina su tarifa, brindando confianza y transparencia en el proceso de tarificación. Esto es fundamental para fomentar una relación sólida y duradera con los asegurados, así como para mantener la competitividad y el crecimiento en el mercado de seguros oncológicos.

Al implementar este modelo analítico, UdeA Insurance se posiciona estratégicamente para tomar decisiones basadas en datos sólidos y evidencia objetiva. La información recopilada y analizada permitirá una tarificación más precisa y justa, asegurando que los asegurados paguen primas que se ajusten a sus perfiles individuales. Esto a su vez ayuda a minimizar los riesgos financieros para la compañía y maximizar las oportunidades de crecimiento.

b. Análisis exploratorio y limpieza de datos

Para el análisis exploratorio se cuenta con 6 bases de datos, las cuales son nombradas de la siguiente manera: {“reclamaciones” - “diagnósticos” - “regional” - “género” - “sociodemográficas” - “utilizaciones”}

Se revisa cada base de datos para comprender la construcción y composición de la información que contienen. Hallando que en las bases de reclamaciones y diagnósticos se cuenta con información general que especifica qué tipo de enfermedades y servicios son prestados, se cuenta con un código asociado que será de utilidad en etapas posteriores, por último se evalúa los datos nulos y las categorías duplicadas, donde se realiza un tratamiento de datos para agrupar en la base de diagnósticos aquellas categorías que no presentan información relevante como lo son datos sin información y diagnósticos pendientes.

También, se exploran las bases de datos regional y género que tienen como función contar con bases aisladas para contener las posibles categorías asociadas al campo de información, encontrando que en regiones existen cinco posibles categorías y en género dos categorías, además que cada una cuenta con una categoría adicional si no se tiene información.

Por otra parte, la base sociodemográfica contiene la información relacionada a cada afiliado, donde se suministra un id y se especifica su género, fecha de nacimiento, la región a la que accede y si posee determinadas enfermedades de base (5 columnas dummies). En esta base se requiere tratar la columna que almacena la información de la fecha de nacimiento, utilizando el formato Excel se converge el dato para obtener la fecha en formato datetime y está a su vez es tratada para obtener la edad en años que es uno de los datos más representativos. Seguido, en esta base de datos se usa para asociar la información de las bases de datos regional y género, pues hasta ese momento la base sociodemográfica cuenta con la codificación y no la categoría.

Finalmente, se explora la base de datos de utilizaciones que brinda información respecto a eventos ocurridos donde se incurra en reclamaciones, el tipo de diagnóstico y la información del id del afiliado, esta base se usa como la principal para unificar las bases y construir una base final que permita hacer una exploración más profunda y completa para la asociación entre variables. También se cuenta con dos campos que hacen referencia a la cantidad de servicios usados en una reclamación y el precio asociado.

Después de explorar cada base de datos, el resultado final es una base de datos consolidada que suministre de forma detallada cada evento y con todas las posibles variables explicativas que tengan las bases de datos. Se verifica la existencia de datos nulos, ya que al unificar bases de datos es posible que en ciertas bases no existan los registros y queden vacíos. De acuerdo a lo anterior se encuentra que existen 4877 datos nulos en diversas columnas, se verifica estas columnas y se encuentra que los datos nulos que registra cada variable explicativa hacen parte de los mimos id para cada columna con lo cuál se considera conveniente eliminar estos registro, no se expone a pérdida de información con lo que no se requiere adoptar técnicas para rellenar campos.

De acuerdo con el problema de negocio, que se asocia concretamente a la tarifación se creó una nueva columna que indicará el precio promedio por reclamación, evitando así el sesgo en la interpretación de datos. Una vez con la base de datos totalmente construida se pasa a hacer un análisis detallado y minucioso de la relación de las variables explicativas con la variable respuesta (objetivo).

Se evalúa cuales reclamaciones tienen costos asociados más elevados, identificando que las reclamaciones por cáncer tienden a ser una de las más altas, además se evalúan las características inherentes a la persona para visualizar si influyen la variable respuesta de manera significativa. Bajo este esquema se encamina el problema de negocio que UdeA Insurance ha decido tratar.

Se acota la base de datos a las reclamaciones por cáncer, dándole el nombre a todas bajo una misma categoría y se realiza un agrupamiento por edades en grupos que comprendan la información de la siguiente manera:

Grupo 1: (0-30], Grupo 2: (30-60], Grupo 3 (60-90] , Grupo 4: (90,105].

Para concluir este apartado, se debe mencionar la transformación de variables categóricas a variables numéricas (género, regiones) por medio del método dummies y la eliminación de los campos de diagnóstico y reclamaciones que dejan de ser relevantes.

c. Selección de algoritmos y técnicas de modelado - Sistemas de recomendación

De acuerdo con la naturaleza de los los datos contenidos en la base, se comprende que sus variables son contínuas y determinando la cantidad y finalidad, se puede definir el modelo como de regresión.

Los algoritmos elegidos fueron:

- Árbol de decisión (**DecisionTree**): Los árboles de decisión son modelos no lineales que dividen el espacio de características en regiones más pequeñas y homogéneas. Estos modelos tienen la capacidad de capturar interacciones complejas entre las variables y son aptos para trabajar con variables numéricas continuas. Sin embargo, es importante tener en cuenta que los árboles de decisión tienden a generar sobreajuste en conjuntos de datos masivos. Esto significa que el modelo puede ajustarse demasiado a los datos de entrenamiento y tener dificultades para generalizar correctamente a nuevos datos.
- Regresión lineal (**LinearRegression**): La regresión lineal es un modelo clásico y ampliamente utilizado que establece una relación lineal entre las variables de salida (objetivo) continua y las variables de entrada. Su objetivo es proporcionar una comprensión clara de la relación y la importancia relativa de cada variable explicativa en la predicción del valor objetivo. Al analizar los coeficientes asociados a cada variable en el modelo de regresión lineal, podemos determinar el impacto que tienen en la predicción final. Esto permite identificar las variables más influyentes y comprender cómo afectan al resultado.

- Modelos de bosques aleatorios (**RandomForests**): Los bosques aleatorios son un enfoque de ensamble que combina múltiples árboles de decisión para obtener una respuesta más robusta y precisa. Este modelo es especialmente útil cuando se trabaja con variables numéricas continuas. La principal ventaja de los bosques aleatorios es su capacidad para proporcionar información sobre la importancia relativa de las variables. Utilizando medidas como la disminución de la varianza en cada ramificación del árbol, podemos identificar qué variables tienen un mayor impacto en la predicción de la variable objetivo. Esto nos permite centrar nuestros esfuerzos en las variables más relevantes y tomar decisiones más fundamentadas.

d. Selección de variables

En la selección de variables se identificó que la gran parte de la composición de las base de datos agregan valor exceptuando las variables asociadas al género. Para realizar la selección se utilizó el método threshold el cuál es un parámetro opcional que controla el umbral de selección de variables. En este caso, se está utilizando el valor de la media, con lo cuál se seleccionan todas las variables cuya importancia sea mayor o igual a la media de la importancia de todas las variables.

e. Evaluación y selección del modelo

Los métodos de evaluación de rendimiento de regresión utilizados para la selección del modelo fueron:

MAE: este calcula las diferencias absolutas entre las predicciones y los valores reales. Se puede interpretar de acuerdo a si es menor, el modelo tiene un mejor desempeño.

RMSE: calcula la raíz cuadrada del promedio de los errores al cuadrado entre las predicciones y los valores reales. De la misma manera en la que se interpreta el MAE, esta métrica indica que entre menor sea el número, mejor es su desempeño.

Se realiza una análisis de máximos y mínimos en la sección *Base_final* con el fin de determinar e interpretar de mejor forma los resultados obtenidos por las métricas de desempeño.

Modelos \ Métricas de desempeño	MAE	RMSE
Árbol de decisión	406857.37670	881912.82366
Regresión Linear	405690.12378	877549.63041
Modelos de bosques aleatorios	406755.42305	881912.82366

De acuerdo con los resultados expuestos en la tabla, los mejores resultados basados en las métricas de desempeño MAE y RMSE son: (405690.12378 y 877549.63041) respectivamente para el modelo de Regresión Linear.

f. Despliegue del modelo

El objetivo final del modelo es apoyar en la toma de decisiones para la tarificación del seguro oncológico, para este proceso el modelo capta la información por medio de un formulario sobre la persona que quiere adquirir el seguro, de acuerdo con las características recopiladas el modelo entregará al departamento de actuaría una base de datos en excel con el id de la persona interesada y el precio que ha determinado el modelo. Este valor es una base que le permitirá al departamento de actuaría evaluar con los criterio de frecuencia y severidad tarifas justas y que mantengan la competitividad de la empresa.

La base de datos será actualizada cada que una persona registre su información y su depuración dependerá de la agilidad del departamento de actuaría en culminar el proceso de venta, aquellas personas que tomen el servicio deben ser agregadas a la base de datos sociodemográficas y cuando estas hagan uso del seguro se registrará la información basados en datos de utilizaciones con la cuál se entrenará el modelo de manera mensual para mejorar sus predicciones.

Por otra parte, se debe tener en cuenta que el modelo no tiene en cuenta las personas que han tenido diagnósticos de cáncer en su historial clínico antes de ingresar a la aseguradora, en caso de aceptar personas con este tipo de novedad se debe tener en cuenta para actualizar el modelo.

g. Conclusiones

- Comprender la cantidad de elementos que pueden componer la cifra que culmina con la tarifación de un seguro, es importante para una empresa, porque de ello depende la generación de una prima justa, y un márgen de ganancia competitivo con el mercado.
- El enfoque que se le pueda dar a los datos, pueden generar propuestas interesantes para ampliar el portafolio de servicios, de acuerdo con el análisis de datos, puede visualizarse eventos crecientes en la población que pueden ser objeto de seguro.

Bibliografía:

- **UNAB.** (2022). *Repositorio de los registros poblacionales de cáncer en Colombia actividades y necesidades orientadas al fortalecimiento de registros poblacionales de cáncer en Colombia vigencia 2022 Convenio 0250.* Recuperado de:

https://apolo.unab.edu.co/es/projects/repositorio-de-los-registros-poblacionales-de-c%C3%A1ncer-en-colombia-

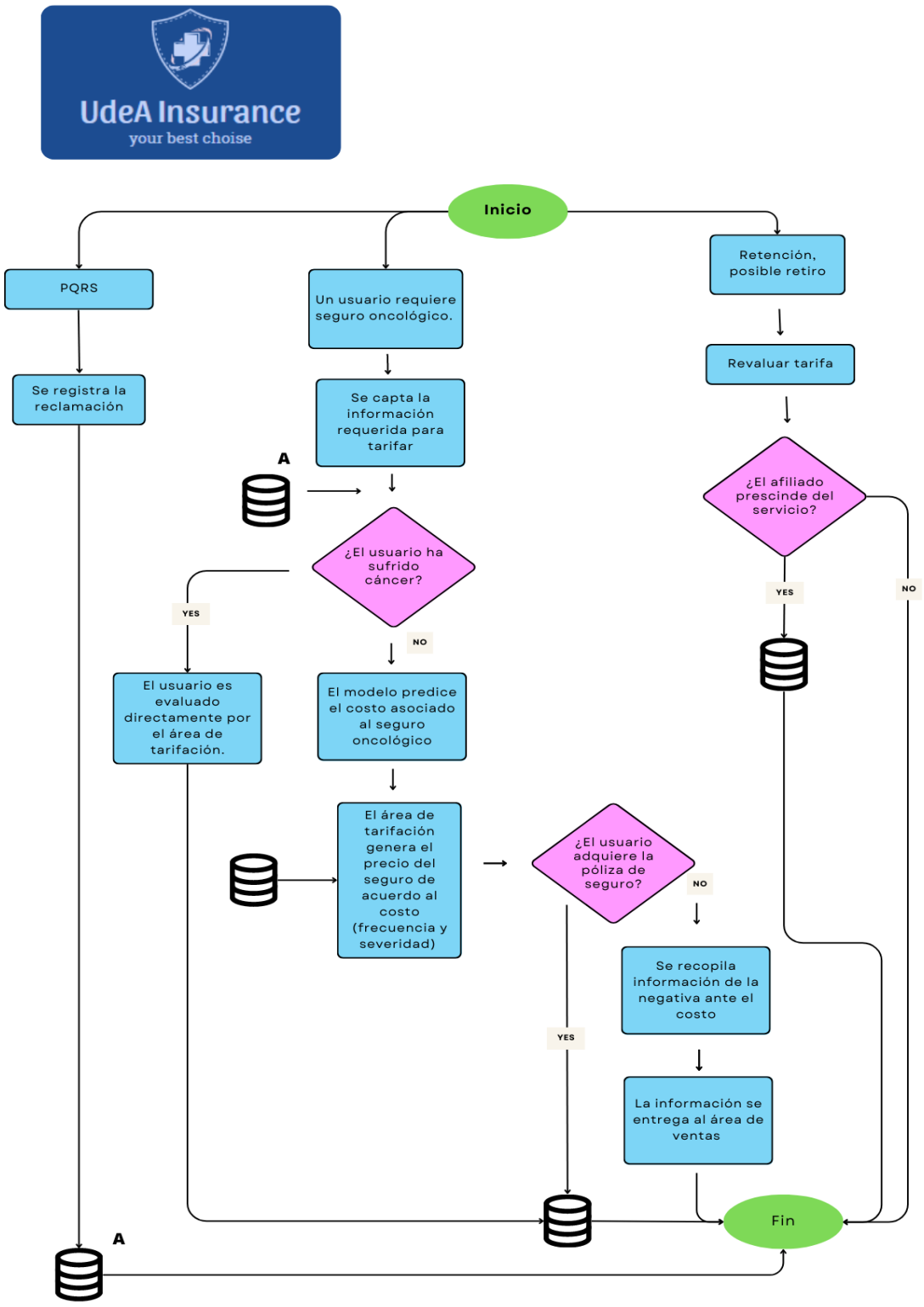
- **ScienceDirect.** (2023). *International Journal of Information Management Data Insights.* Recuperado de:

https://www.sciencedirect.com/journal/international-journal-of-information-management-data-insights

- **BBVA. (2018)** *Póliza de Seguro Oncológico. Recuperado de:*

https://www.bbva.com.co/personas/productos/seguros/libres/oncologico.html#mas-informacion-sobre-coberturas

Anexo:


UdeA Insurance
your best choise
Inicio
PQRS
Un usuario requiere seguro oncológico.
Retención, posible retiro
Se registra la reclamación
Se capta la información requerida para tarifar
Revaluar tarifa
A
¿El afiliado prescinde del servicio?
¿El usuario ha sufrido cáncer?
YES
NO
YES
NO
El usuario es evaluado directamente por el área de tarifación.
El modelo predice el costo asociado al seguro oncológico
El área de tarifación genera el precio del seguro de acuerdo al costo (frecuencia y severidad)
¿El usuario adquiere la póliza de seguro?
NO
YES
Se recopila información de la negativa ante el costo
La información se entrega al área de ventas
Fin
A