Universidad de Santiago de Chile

Facultad de Administración y Economía

Diplomado Data Science

Generación de un modelo de estimación para determinar la adjudicación de presupuesto CORFO a una empresa según sus características

Amilcar Rodriguez, R.U.T.: 25.868.813-9

Santiago de Chile, enero 2024

Índice

1. Introducción 3
2. Correcciones implementadas 3
2.1. Comprensión de los datos 3
2.2. Preparación de los datos 4
3. Modelamiento 5
3.1. Selección de aproximaciones a la variable dependiente 5
3.2. Definición de pruebas 6
3.3. Diseño y aplicación 6
4. Evaluación de los resultados 6
4.1. Comparación de rendimientos 6
4.2. Revisión del flujo del proceso. 10
4.3. Determinar próximas etapas 11
5. Anexos 12
5.1. Archivo ejecutado Jupiter del proyecto 12

1. Introducción

En base a los dos entregables anteriores que abordaron las fases de:

- Comprensión del negocio
- Comprensión de los datos
- Preparación de los datos

Se ha hecho una revisión del proyecto y por ende se han implementado una serie de correcciones que buscan sincronizar la preparación de los datos y comprensión del negocio con las fases posteriores de modelamiento y evaluación de resultados, para de esta forma cumplir con las distintas etapas del flujo análisis de los datos y más importante aún dar respuesta a las preguntas de valor y objetivos del negocio impuestos al inicio del proyecto.

A continuación, se desarrollan y fundamentan algunas de las correcciones implementadas como también las etapas de modelamiento y evaluación.

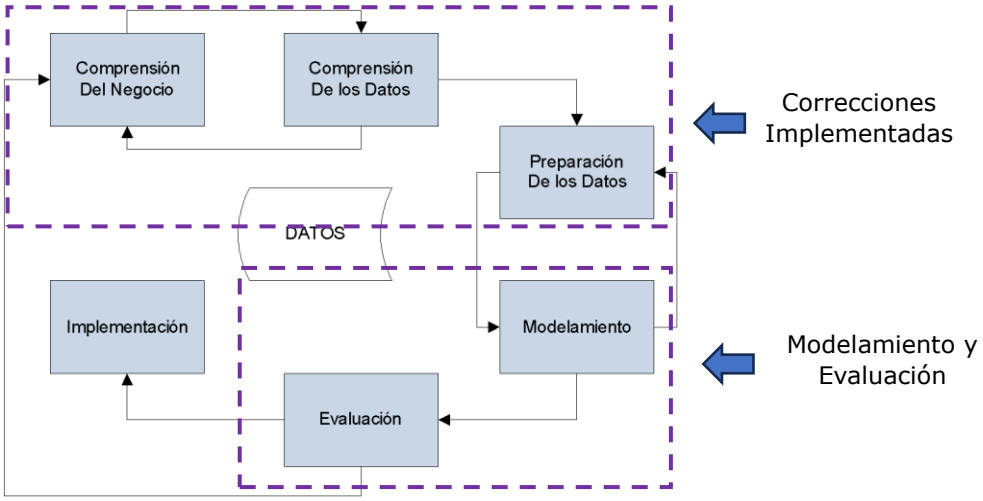

Imagen 1. Flujo de proyecto y alcances finales del proyecto

2. Correcciones implementadas

2.1. Comprensión de los datos

A nivel de Comprensión de los datos, se han realizado un nuevo análisis para mejorar el entendimiento del negocio y del comportamiento de los datos, para esto se ha realizado un nuevo análisis exploratorio buscando validar e identificar nuevamente las características que serían las entradas para las fases de posteriores del proyecto.

Sobre la misma línea de ideas existe un cambio importante que se da desde la adquisición de los datos, y es que se ha cambiado la fuente de datos para el data set de CORFO – Data Innova, anteriormente estos datos se descargaban manualmente desde el link https://datainnovacion.cl/portafolio-proyectos, sin embargo, se han hecho cambios en el código para leer la información vía API y Python.

La principal motivación para actualizar la fuente de datos para CORFO – Data Innova, es que en la muestra anterior de datos (Entrega 02), todos los casos del portafolio, eran casos de presupuestos "Aprobado" CORFO, los cuales se han asumido utilizando la característica "aprobado_corfo" como mayor a 0, pero no existían casos iguales a 0, por lo tanto, en vista de esta carencia, que se asume importante para la predicción y clasificación de casos, se buscó acceder a la data completa donde si existan registros de montos a 0, los cuales podríamos asumir como "No aprobados".

Al hacer el cruce de los datos, es decir, ambos datasets, tanto CORFO como SII, y evaluando el porcentaje de casos de presupuesto aprobados y no aprobados, se tiene la siguiente distribución porcentual:

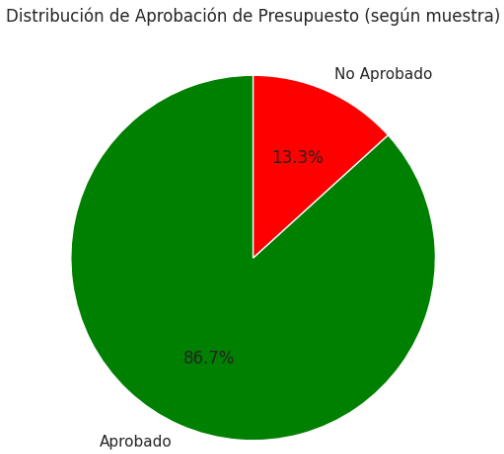

Imagen 2. Porcentajes Casos Aprobados y No Aprobados

2.2. Preparación de los datos

La variable objetivo de este proyecto se ha identificado como "aprobado_corfo" naturalmente existe dentro del *dataframe* en un formato numérico, sin embargo, es requerido agregar el enfoque discreto al análisis debido a la naturaleza de los objetivos de negocio, que buscan identificar las variables claves para la adjudicación del presupuesto, por lo tanto, haciendo referencia a la variable "aprobado_corfo" se crea la variable categórica "estado_aprobacion" que contiene a su vez las opciones "Aprobado" ("aprobado_corfo" > 0) y "No Aprobado" ("aprobado_corfo" = 0)

En cuanto a la preparación de los datos (en su conjunto), ya sobre un dataframe consolidado con escenarios que permitan poder hacer clasificaciones y predicciones, se realizó una nueva exploración de los datos, y se revisó y validó la selección de características. Sobre esta nueva exploración de los datos, se profundizó el análisis y procesamiento de variables categóricas como:

- "Razon_social"
- "Rubro_economico"
- "Actividad económica"
- "Región"
- "tendencia final"

- "mercado_objetivo_final"

Las cuales demostraron ser claves para la variable dependiente pero que además contienen un alto número de categorías nominales, esto hizo necesario aplicar mapeos de variables a mano en base al contexto de negocio y algunos algoritmos de *text mining* y aprendizaje no supervisado como k-means para la simplificación de características.

En concreto las características finalmente seleccionadas para la etapa de modelamiento, se tienen:

Tabla 1. Características seleccionadas para el modelamiento

#	Característica	Tipo de dato
0	aprobado_privado	Int64
1	aprobado_privado_pecuniario	Int64
2	Numero_de_trabajadores_dependientes	Int64
3	**aprobado_corfo (variable dependiente continua)**	Int64
4	tipo_persona_beneficiario	Object
5	tramo_ventas	Object
6	Subtipo_de_contribuyente	Object
7	Rubro_economico	Object
8	Región	Object
9	tendencia_final	Object
10	mercado_objetivo_final	Object
11	tipo_innovacion	Object
12	Tipo_de_contribuyente	Object
13	tipo_proyecto	Object
14	sostenible	Object
15	economia_circular_si_no	Object
16	genero_director	Object
17	ley_rep_si_no	Object
18	**estado_aprobacion (variable dependiente discreta)**	Object
19	razon_social_agrupadas	Object

3. Modelamiento

3.1. Selección de aproximaciones a la variable dependiente

Las aproximaciones que se han realizado son las siguientes:

Tabla 2. Aproximaciones realizadas para el modelamiento

		Tipo de Algoritmo	
		Agnóstico	Estadístico
Aproximación	Continua	Support Vector Machines (SVM)	Regresión Lineal
	Discreta	Arboles de Decisión	Regresión Logística

3.2. Definición de pruebas

Las métricas que se han seleccionado son las siguientes por aproximación:

- Continua

  (1) Coeficiente de determinación (R^2).
  (2) Error cuadrático medio (RMS).

Estas métricas se han seleccionado debido a su capacidad de medir la proporción de la varianza en los datos respecto a las características seleccionadas (R^2), de esa manera podríamos medir que tan bien están representando los datos la realidad del objetivo buscado y por otro lado evaluar precisión (RMS) que para efectos del problema podría verse impactado por el tipo de problema que se buscar resolver de fondo (clasificación o predicción)

- Discreta

  (1) Accuracy.
  (2) Recall.
  (3) F1-Score.
  (4) Matriz de Confusión.

Estas métricas se han seleccionado debido a que es posible con ellas medir precisión (*Accuracy*), la sensibilidad (*Recall*) para determinar verdaderos positivos y obtener una precisión robusta, también se agrega la matriz de confusión para obtener una visión más completa del rendimiento del modelo y de los posibles resultados.

3.3. Diseño y aplicación

La parametrización de los datos se ha hecho en base a un 20% de datos para test y un 80% datos de entrenamiento, se han hecho pruebas con el parámetro de *stratify* para manejar el desbalance de clases, el cual es claro y ha quedado manifestado en el gráfico de la imagen 1, sin embargo, al haber utilizado este parámetro no se han obtenido resultados significativos, esto se debe a los resultados de los modelos de clasificación que a continuación se verán.

4. Evaluación de los resultados

4.1. Comparación de rendimientos

A continuación, se comparan los resultados por cada tipo de aproximación

- Continuas

Tabla 3. Comparación de métricas de rendimiento Aproximaciones Continuas

Aproximaciones Continuas	
Support Vector Machines (SVM)	**Regresión Lineal**
MSE: 2513014670302720.0 R^2: 0.5521052302982833	MSE: 6040729502172142.0 R^2: -0.07663961582836154
• **Mean Squared Error (MSE):** este valor es bastante alto (2.51e15), lo cual sugiere que las predicciones del modelo están bastante alejadas de los valores reales. Cuanto menor sea el MSE, mejor, ya que indica un mejor ajuste del modelo a los datos. • **R^2 Score:** el valor de R^2 es 0.55, lo cual indica que el modelo explica aproximadamente el 55% de la variabilidad en los datos. Este valor es positivo, lo cual es un buen signo, pero podría sugerir que hay margen para mejorar la capacidad del modelo para explicar la variabilidad.	• **Mean Squared Error (MSE):** el MSE es extremadamente alto (6.04e15), lo cual indica que las predicciones del modelo están muy lejos de los valores reales. Este alto error puede deberse a diversos motivos, como la sensibilidad a los hiperparámetros o la necesidad de escalar las características. • **R^2 Score:** El R^2 Score es negativo (-0.07), lo cual indica que el modelo no está realizando una buena predicción y es incluso peor que un modelo constante que siempre predice la media.

Estos resultados si bien no son buenos, pueden tener mucho sentido y se debe al propósito y objetivos del proyecto, ya que no se busca predecir el monto o presupuesto aprobado por CORFO para cado caso, si no, se busca clasificar y entender el comportamiento de las características sobre una variable binaria discreta (Aprobado, No Aprobado). Por otra parte, podríamos complementar y recomendar:

A nivel de preprocesamiento: se podría considerar la opción de evaluar las escalas de los datos y buscar escalarlos en caso de ser necesario para la aplicación de SVM.

A nivel de la selección de las variables: se han realizados distintas iteraciones entre la entrega 2 y esta entrega (3) en lo que respecta al análisis y selección de los datos, sin embargo, se podría evaluar iterando buscando profundizar y entonar más este apartado, principalmente en la selección y preprocesamiento de variables como Razón Social, Actividad económica, etc.

Validación Cruzada (CV): se recomienda utiliza CV para obtener evaluaciones más robustas del rendimiento del modelo, esto se hizo de manera acotada a nivel de pruebas, no se ha dejado documentado en este archivo debido a que los resultados no fueron muy distintos, sin embargo, considerando las otras recomendaciones podría ser atractivo repetir este tipo de acción.

Ajustar Hiperparámetros: en el caso de SVM la elección correcta de los hiperparámetros es crucial. se recomienda probar con diferentes configuraciones y utiliza técnicas específicas para la búsqueda de mejores parámetros, sin embargo, esto se ha decidido dejar de lado para este caso debido a la naturaleza de la variable objetivo, que es fundamentalmente categórica debido al problema que se busca resolver.

Modelos Alternativos: se recomienda utilizar modelos de regresión como Lasso o Ridge o incluso modelos de conjunto como *Gradient Boosting* ya que estos pueden ser más robustos y ayudar a mejorar los "resultados".

- Discretas

Tabla 4. Comparación de métricas de rendimiento Aproximaciones Discretas

Aproximaciones Discretas
Regresión logística
• **Accuracy:** 0.9980353634577603 • **Reporte de Clasificación:** **precision recall f1-score support** **Aprobado** 1.00 1.00 1.00 1310 **No Aprobado** 0.99 1.00 0.99 217 **accuracy** 1.00 1527 **macro avg** 0.99 1.00 1.00 1527 **weighted avg** 1.00 1.00 1.00 1527 • **Matriz de Confusión:** [[1307 3] [0 217]]
• **Accuracy:** La precisión general del modelo es del 99.80%, lo cual indica que el modelo está clasificando correctamente la gran mayoría de las instancias. • **Reporte de Clasificación:** Las métricas de precisión, recall y f1-score son muy altas para ambas clases (0 y 1). Estos valores cercanos a 1 sugieren un rendimiento muy bueno del modelo. • **Matriz de Confusión:** La matriz de confusión muestra que el modelo cometió solo 3 falsos positivos y ningún falso negativo. Esto es una señal de un modelo muy preciso.
Arboles de Decisión
• **Accuracy:** 1.0 • **Reporte de Clasificación:** **precision recall f1-score support** **Aprobado** 1.00 1.00 1.00 1310 **No Aprobado** 1.00 1.00 1.00 217 **accuracy** 1.00 1527 **macro avg** 1.00 1.00 1.00 1527 **weighted avg** 1.00 1.00 1.00 1527 • **Matriz de Confusión:**

[[1310 0] [0 217]]
• **Accuracy:** La precisión general del modelo es aproximadamente de 1. • **Reporte de Clasificación:** Las métricas de precisión, recall y f1-score son muy altas, mostrando que el modelo está logrando una buena clasificación para ambas clases. • **Matriz de Confusión:** Similar al modelo de Regresión Logística, el Árbol de Decisión muestra un rendimiento muy bueno, sin falsos positivos y ningún falso negativo.

El rendimiento de los modelos aparentemente es prometedor, por lo que podríamos concluir y definir las variables claves que influyen en la toma de decisiones, las cuales son en base a un umbral de 0.5 (moderado) calculado por los coeficientes de la regresión logística:

Tabla 5. Características y clases con mayor influencia sobre las predicciones (con umbral)

index	Característica	Coeficiente
3	tipo_persona_beneficiario_PERSONA JURIDICA COMERCIAL	2.6107669420731100
4	tipo_persona_beneficiario_Persona Jurídica constituida en Chile	2.3045116840226800
10	tramo_ventas_Sin ventas	-0.5696657942297434
68	mercado_objetivo_final_Educación y Servicios Conexos	-0.5024779610895707
81	mercado_objetivo_final_Multisectorial	-0.5920911036691782
108	razon_social_agrupadas_Desarrollo Agropecuario e Institucional	0.7673640661623546

Sin embargo, todas las características (Tabla 1) parecen tener un buen rendimiento (con mayor o menor peso) y pesar sobre la variable dependiente. Podríamos complementar y recomendar:

Desbalance de Clases: se observa que hay un desbalance en las clases, ya que la clase 0 (Aprobado) tiene muchos más ejemplos que la clase 1 (No Aprobado).

Validación Cruzada: aunque los resultados aparentemente son buenos, podría es útil aplicar validación cruzada para obtener una validación más robusta de los modelos.

Modelos Alternativos: si el desbalance de clases se convierte en un problema, se podría explorar submuestreo (*undersampling*) o sobremuestreo (*oversampling*), o probar con modelos de conjuntos

En general, los resultados actuales indican un rendimiento muy bueno de los modelos para este tipo de aproximación, de todas maneras, esto podría ser "sospechoso" y no se

deben malinterpretar los resultados, se recomienda seguir profundizando e iterando en base a las recomendaciones anteriormente mencionadas.

4.2. Revisión del flujo del proceso.

Dentro del flujo o proceso se han presentado distintos dificultades y desafíos, así como también se han propuesto soluciones dentro del mismo proyecto, algunas un poco más sencillas como la creación de una variable categórica (objetivo) en base a una numérica y otras más generales y teóricas, también se ha buscado trabajar sobre una línea fundamentada en los conceptos aprendidos durante el diplomado. A continuación, se enlistan algunos desafíos y complejidades como también acciones resolutivas o puntos de mejora identificables para cada uno.

- Cantidad de datos disponibles. La cantidad de datos deja abierta la puerta para seguir analizando los modelos propuestos, debido a que existe un desbalance de clases que pudiera estar afectando los resultados de las estimaciones. Se podría considerar la utilización de técnicas adicionales para refinar los modelos y resultados.

- Alto volumen de categorías por características y presencia de multicolinealidad. Esto ha representado un desafío ya que como parte del preprocesamiento de los datos lo que se pretende es refinar, mantener y mejorar la calidad de estos, no lo contrario, sin embargo, se ha buscado trabajar con el menor número de variables posibles para la simplicidad del trabajo y esto ha conllevado a utilizar algunos métodos para procesar y simplificar las características.

- Experiencia y conocimiento del negocio para el análisis y abordaje del proyecto. Debido a que el tema del proyecto de título no es un tema donde inicialmente este equipo de trabajo (mi persona) tenga particular experiencia profesional, el inicio del flujo ha representado un desafío, sin embargo, gracias a el proceso (iterativo) definido y sus diferentes etapas, la temática ha sido digerida e interpretada para los posteriores avances.

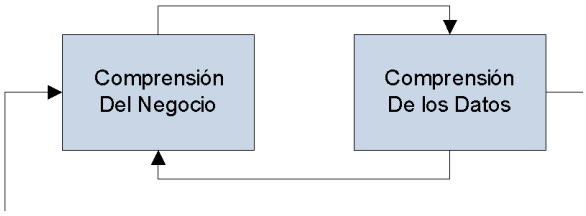

Imagen 3. Etapa para la comprensión del negocio y datos

- Automatización y codificación de la solución. Un alto porcentaje del trabajo se ha dado bajo la digestión y preparación de los datos, sin embargo, esta actividad consume mucho tiempo debido a los distintos análisis y enfoques que se pueden aplicar sobre los mismos, por lo que pudiera recomendarse para etapas futuras la automatización de al menos las acciones de análisis bases, para así reducir los tiempos y plazos de análisis y definición de estrategias.

4.3. Determinar próximas etapas

Este proyecto ha sido abordado con propósitos académicos, sin embargo, es un flujo que podría seguirse trabajando, mejorando y corrigiendo, ya que, en base a lo anterior, el que los resultados de los modelos de clasificación sean buenos, no quiere decir que se necesiten hacer más pruebas para la validación de los resultados, sin embargo, en alcance definido se han podido identificar características claves que pesan en la toma decisiones para la aprobación de presupuestos.

5. Anexos

5.1. Archivo ejecutado Júpiter del proyecto

Proyecto Título -
Entrega 03 - Amilcar