Universidad Nacional Autónoma de México

IIMAS

Introducción a la Empresa y a las Finanzas

Semestre 2025-1

Proyecto Final

Escrito por:

Lagunas Parra Jaime Rodrigo

Lara Nieva Allison

Salgado Tirado Diana Laura

1. Introducción

En el presente proyecto nos enfocamos en el análisis de indicadores, el cuál es fundamental para evaluar el desempeño y la estabilidad de las empresas, así como para identificar factores clave que puedan influir en sus costos de capital y su rentabilidad. Específicamente examinaremos las relaciones entre *Cost of Equity (CoE)*, que refleja el rendimiento requerido por los inversionistas para compensar el riesgo de invertir en una empresa, y un conjunto de variables independientes que representan características financieras y operativas de 32 empresas.

La importancia de comprender la relación entre el *CoE* y los diferentes indicadores con los que contamos, permitiría a inversionistas, analistas o personas con el poder de tomar decisiones de identificar los principales determinantes del costo de financiamiento.

Metodología

La estructura del proyecto tendrá 2 enfoques principales (lineal y no lineal). En primer lugar, se emplearán diagramas de dispersión para visualizar la relación entre el *CoE* y las variables independientes. Lo que nos permite identificar patrones, tendencias y posibles correlaciones, ya sea positivas, negativas o inexistentes, entre las variables, y así interpretar fácilmente los datos con los que contamos.

Enfoque Lineal

Para el enfoque lineal, se utilizará el cálculo de la correlación entre las variables independientes y la variable objetivo *CoE*. Esta correlación se mide mediante el coeficiente de correlación de Pearson, el cual se calcula utilizando la siguiente fórmula:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \cdot \sum(Y_i - \bar{Y})^2}}$$

Donde:

- X_i y Y_i son los valores individuales de las variables.
- $\bar{X}$ y $\bar{Y}$ son las medias de X y Y, respectivamente.

El coeficiente r toma valores entre -1 y 1, los cuales se interpretan de la siguiente manera:

- $r = 1$: Relación perfectamente positiva.
- $r = -1$: Relación perfectamente negativa.
- $r = 0$: Sin relación lineal.

Posteriormente, las variables se ordenarán por magnitud según el valor absoluto de su correlación, priorizando aquellas con mayor relación (positiva o negativa) con la variable objetivo. Este procedimiento permite identificar las características que aportan más información relevante sobre el *CoE*.

Es importante destacar que este método solo mide relaciones lineales entre las variables. Si la relación entre una característica y la variable objetivo es no lineal, el coeficiente de Pearson podría no reflejar adecuadamente su impacto, ya que este enfoque analiza exclusivamente la relación bivariada y no contempla interacciones entre múltiples variables.

Enfoque No Lineal

Para el enfoque no lineal, se utilizará un modelo de regresión basado en *Random Forest*, donde el criterio principal para la evaluación de las divisiones en los nodos es el *Error Cuadrático Medio* (*Mean Squared Error*, MSE). Este criterio se fundamenta en minimizar la varianza dentro de los nodos resultantes después de una división, buscando obtener subconjuntos más homogéneos.

La medida de impureza para regresión en un nodo N con un conjunto de datos S se calcula como la varianza de los valores objetivo y:

$$\text{Varianza}(N) = \frac{1}{|S|} \sum_{i \in S} (y_i - \bar{y})^2$$

Donde:

- y_i: Valor objetivo de la instancia i.
- $\bar{y}$: Promedio de los valores objetivo en S.
- $|S|$: Número de instancias en S.

Cuando el nodo N se divide en dos subconjuntos, S_L y S_R, el *MSE* evalúa la calidad de la división como:

$$\text{MSE} = \frac{|S_L|}{|S|} \cdot \text{Varianza}(S_L) + \frac{|S_R|}{|S|} \cdot \text{Varianza}(S_R)$$

La división óptima es aquella que minimiza este valor de *MSE*, lo que implica reducir la varianza dentro de los subconjuntos resultantes.

En el modelo de *Random Forest Regressor*, la importancia de una característica X_i se calcula evaluando cuánto contribuye esta a reducir la varianza (*MSE*) en los nodos donde es utilizada como criterio de división. Matemáticamente, se define como:

$$\text{Importancia}(X_i) = \sum_{t \in \text{nodos donde } X_i \text{ es usado}} \frac{|S_t|}{|S|} \Delta\text{Varianza}_t$$

Donde:

- S_t: Tamaño del subconjunto de datos en el nodo t.
- $|S|$: Tamaño total del conjunto de datos.

- $\Delta\text{Varianza}_t$: Reducción de la varianza en el nodo t al utilizar X_i como criterio de división.

Este cálculo se realiza para cada característica, acumulando la reducción de la varianza en todos los nodos donde se utiliza. Las características con valores más altos de importancia tienen un mayor impacto en la predicción del *Cost of Equity (CoE)*, ya que contribuyen de manera significativa a reducir la impureza en el modelo. Esto proporciona una métrica clara y cuantitativa sobre la relevancia de cada variable dentro del modelo no lineal.

Objetivos

- Determinar los principales determinantes del *Cost of Equity (CoE)* mediante el análisis de variables financieras y operativas de 32 empresas.
- Evaluar la relevancia de estas variables utilizando enfoques lineales y no lineales.

Nota: Para un análisis detallado por año, por favor consulte el *notebook* en : `[https://colab.research.google.com/drive/1ZS5eHRHqHLPJ6m4vt_eZbSErsG9b3bbr?usp=sharing]`.

Análisis de los Diagramas de Dispersión

Los siguientes diagramas de dispersión permiten explorar visualmente la relación entre el *Cost of Equity* y diversos indicadores con los que contamos. Se observa que variables como *Ln B/M* y *ROA* presentan patrones más definidos: empresas con mayores valores en estas métricas tienden a tener un menor *CoE*, lo que podemos asociar con menor riesgo percibido tiende a menores costos de capital. Por otro lado, variables como *CARBON*, *Beta* y *OISstd* muestran relaciones menos claras.

Variables como el (*Lev*) sugieren una ligera relación inversa, posiblemente debido a estructuras financieras más optimizadas. *IND*, que parece reflejar un indicador binario. Una vez realizados estos análisis, se justifica un trabajo más profundo, para confirmar la relevancia de las variables y comprender mejor las dinámicas que determinan el costo de capital.

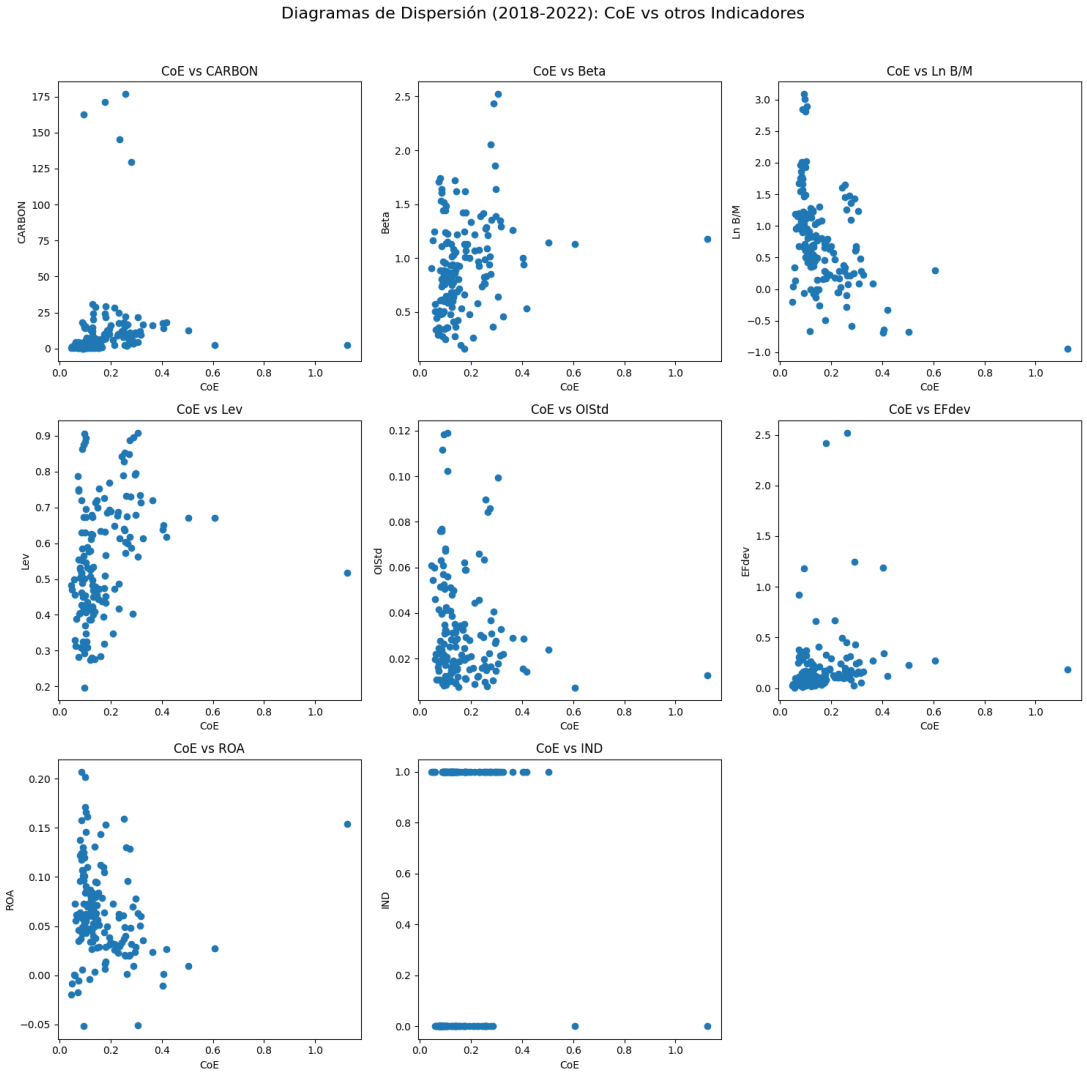

Figura 1: Diagramas Dispersión

Análisis Modelo Lineal

Para analizar la relación entre las variables independientes y el (CoE), se ajustó primero un modelo de regresión lineal utilizando el método de Mínimos Cuadrados Ordinarios (OLS), lo que nospermite identificar las variables más significativas que explican las variaciones en el CoE.

Los coeficientes β se estiman minimizando la Suma de los Cuadrados de los Residuos (RSS):

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}))^2$$

La solución se calcula mediante la fórmula de los Mínimos Cuadrados Ordinarios (OLS):

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Resultados del Modelo Lineal

El modelo ajustado presentó los siguientes resultados clave:

```
R-cuadrado: 0.528
R-cuadrado ajustado: 0.502
```

Esto indica que el 52.8 % de la variabilidad del CoE está explicada por las variables independientes. El valor ($R^2_{ajustado}$) corrige el efecto del número de variables, manteniendo un ajuste moderado del modelo.

Los coeficientes estimados y su significancia son los siguientes:

Variable	Coeficiente	Error Estándar	p-valor
constante	-0.0735	0.035	0.038
CARBON	-0.0005	0.000	0.091
Beta	0.0369	0.017	0.035
Ln B/M	-0.1352	0.013	0.000
Lev	0.4610	0.051	0.000
OIStd	-0.6593	0.303	0.031
EFdev	0.0276	0.023	0.235
ROA	1.2097	0.201	0.000
IND	-0.0053	0.015	0.721

Interpretación de los Resultados

- **Lev ($p < 0.001$):** El apalancamiento financiero tiene una relación positiva con el *CoE*, lo que nos dice que un mayor uso de deuda aumenta el riesgo financiero y, por ende, el costo de capital.
- **Ln B/M ($p < 0.001$):** Las empresas que cuentan con una mayor proporción de valor contable sobre valor de mercado tienden a tener menores costos de capital, lo que podría refleja menor riesgo.
- **ROA ($p < 0.001$):** Empresas más rentables suelen percibirse como menos riesgosas.
- **OIStd ($p < 0.05$):** Las empresas con mayor desviación en sus ingresos operativos presentan un *CoE* más bajo.
- Variables como *CARBON*, *EFdev*, e *IND* no mostraron un efecto significativo.

Evaluación del Desempeño del Modelo Lineal

Para medir la calidad del modelo, calculamos el RMSE Relativo ($RMSE_{relativo}$):

$$RMSE_{relativo} = \frac{\text{RMSE}}{\bar{y}}$$

Donde:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

El RMSE Relativo calculado es 0.4896, lo que indica que el error del modelo es aproximadamente el 49 % del promedio de los valores reales del *CoE*, sugiere que el modelo captura patrones interesantes pero deja espacio a mejora.

Importancia de las Características

Para validar los resultados, se calculó el coeficiente de correlación de Pearson entre las variables independientes y el *CoE*. Las variables más relevantes fueron:

- **Lev:** 0.295 (correlación positiva más fuerte).
- **Ln B/M:** -0.435 (correlación negativa más fuerte).

Los intervalos de confianza para los coeficientes confirman que *Lev*, *Ln B/M*, *ROA*, y *Beta* tienen efectos significativos sobre el *CoE*, ya que sus intervalos no incluyen el 0.

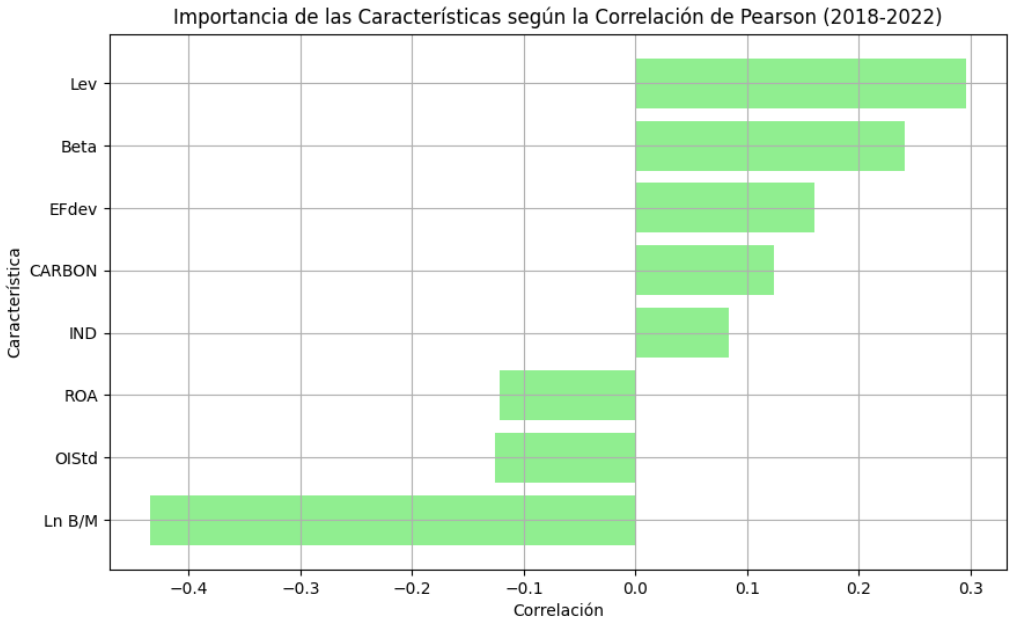

Figura 2: Importancia Características Modelo Lineal

Análisis del Modelo No Lineal: Random Forest

Para el enfoque no lineal se implementó utilizando un modelo de **Random Forest**, que combina múltiples árboles de decisión para mejorar la precisión de las predicciones.

El modelo de Random Forest se basa en la predicción promedio de T árboles de decisión independientes:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x)$$

Donde:

- $h_t(x)$: Predicción del árbol t para la entrada x.
- T: Número total de árboles en el modelo.

Cada árbol se entrena sobre un subconjunto de datos (*bootstrap sampling*) y selecciona características aleatorias para dividir los nodos, minimizando la varianza dentro de cada nodo.

La medida de impureza utilizada para la regresión es la **varianza** de los valores objetivo en un nodo N:

$$\text{Varianza}(N) = \frac{1}{|S|} \sum_{i \in S} (y_i - \bar{y})^2$$

Donde:

- y_i: Valor objetivo de la instancia i.
- $\bar{y}$: Promedio de los valores objetivo en S.
- $|S|$: Número de instancias en el nodo.

El **Error Cuadrático Medio (MSE)** evalúa la calidad de las divisiones, reduciendo la varianza dentro de los subconjuntos resultantes:

$$MSE = \frac{|S_L|}{|S|} \cdot \text{Varianza}(S_L) + \frac{|S_R|}{|S|} \cdot \text{Varianza}(S_R)$$

Donde:

- S_L y S_R: Subconjuntos de datos tras la división.
- $|S|$: Tamaño del nodo original.

Resultados del Modelo

El desempeño del modelo se evaluó utilizando el RMSE y el RMSE relativo:

$$\text{RMSE Relativo} = \frac{\text{RMSE}}{\bar{y}}$$

Los valores obtenidos fueron:

- **RMSE:** 0.0365
- **RMSE Relativo:** 0.2169 (21.69 %)

Esto indica que el modelo tiene un error moderado, capturando patrones significativos en los datos con un error relativo menor al 22 % del promedio de los valores reales.

Importancia de las Características

La importancia de cada característica (X_i) se calcula evaluando cuánto reduce la varianza (MSE) en los nodos donde se utiliza como criterio de división:

$$\text{Importancia}(X_i) = \sum_{t \in \text{nodos donde } X_i \text{ es usado}} \frac{|S_t|}{|S|} \Delta\text{Varianza}_t$$

Las características más importantes en este modelo fueron:

Característica	Importancia
Ln B/M	0.481590
CARBON	0.161834
Lev	0.112474
OIStd	0.105908
Beta	0.053690

Intervalos de Confianza para la Importancia

La importancia se validó mediante un enfoque de permutación, evaluando el impacto en el error al desordenar aleatoriamente una característica. Los intervalos de confianza se calcularon usando **bootstrapping**:

$$\text{Importancia} = \text{Error}_{\text{permutado}} - \text{Error}_{\text{original}}$$

Los resultados fueron los siguientes:

Característica	Media de Importancia	IC Inferior (2.5 %)	IC Superior (97.5 %)
Ln B/M	0.417846	0.136620	0.718944
CARBON	0.177865	0.060056	0.420212
Lev	0.119107	0.039264	0.266488
OIStd	0.114907	0.014572	0.314027
Beta	0.060370	0.021915	0.135286

Interpretación de los Resultados

- **Ln B/M:** Es la variable más importante, lo que refuerza su relación negativa significativa con el *CoE*, como se observó en el modelo lineal.
- **Lev y OIStd:** Tienen un impacto considerable, destacando la relevancia del apalancamiento financiero y la estabilidad de los ingresos operativos.
- **CARBON y Beta:** Aunque menos importantes, contribuyen moderadamente a las predicciones del modelo.

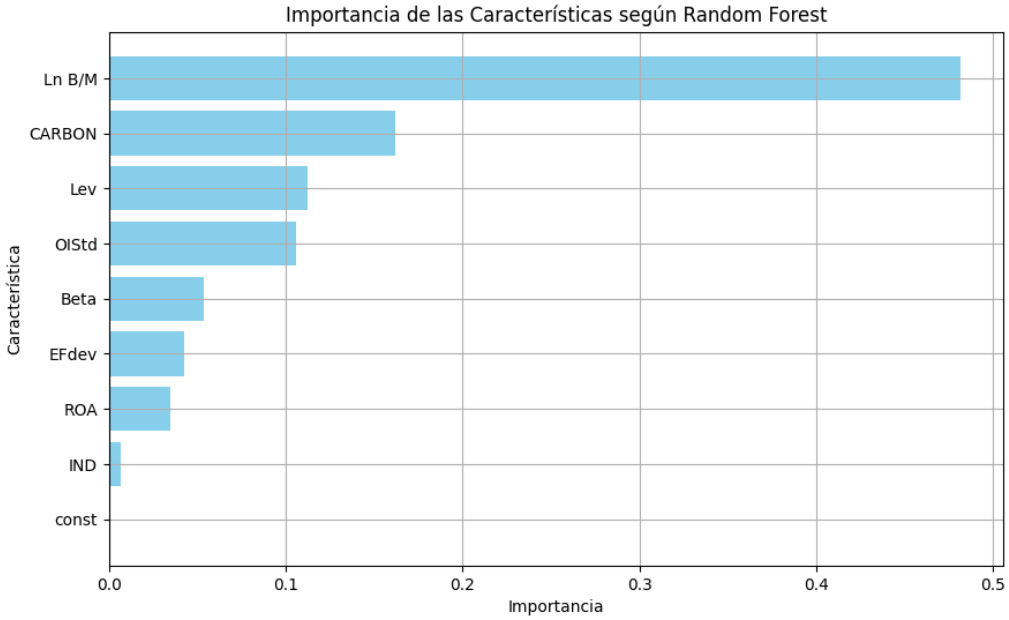

Figura 3: Importancia Características No Lineal

Diferencias entre Enfoques

- **Interpretación:** El modelo lineal es más fácil de interpretar, mientras que el Random Forest identifica patrones más complejos pero es menos intuitivo.
- **Desempeño:** El Random Forest tuvo un menor error relativo (21.69 % vs. 49 %), lo que sugiere que modela mejor las relaciones no lineales en los datos.
- **Variables Relevantes:** Ambos modelos coinciden en la importancia de *Ln B/M* y *Lev*, pero el Random Forest también resalta otras variables como *OIStd*.

Conclusión

En este proyecto se analizaron los factores que influyen en el (CoE) utilizando dos enfoques: un modelo lineal de regresión y un modelo no lineal basado en Random Forest. En el modelo lineal se

identificar relaciones directas y significativas entre las variables, destacando el impacto positivo del (*Lev*) y el efecto negativo del (*Ln B/M*).

En el enfoque no lineal, el modelo de Random Forest mejoró el desempeño predictivo, reduciendo el error relativo al 21.69 %. Además de confirmar la importancia de *Ln B/M* y *Lev*, este modelo resaltó otras características relevantes, como (*OIStd*). Este enfoque es más adecuado para datos complejos. Ambos enfoques, utilizados en conjunto, ofrecen una visión completa en las relaciones del *CoE*.