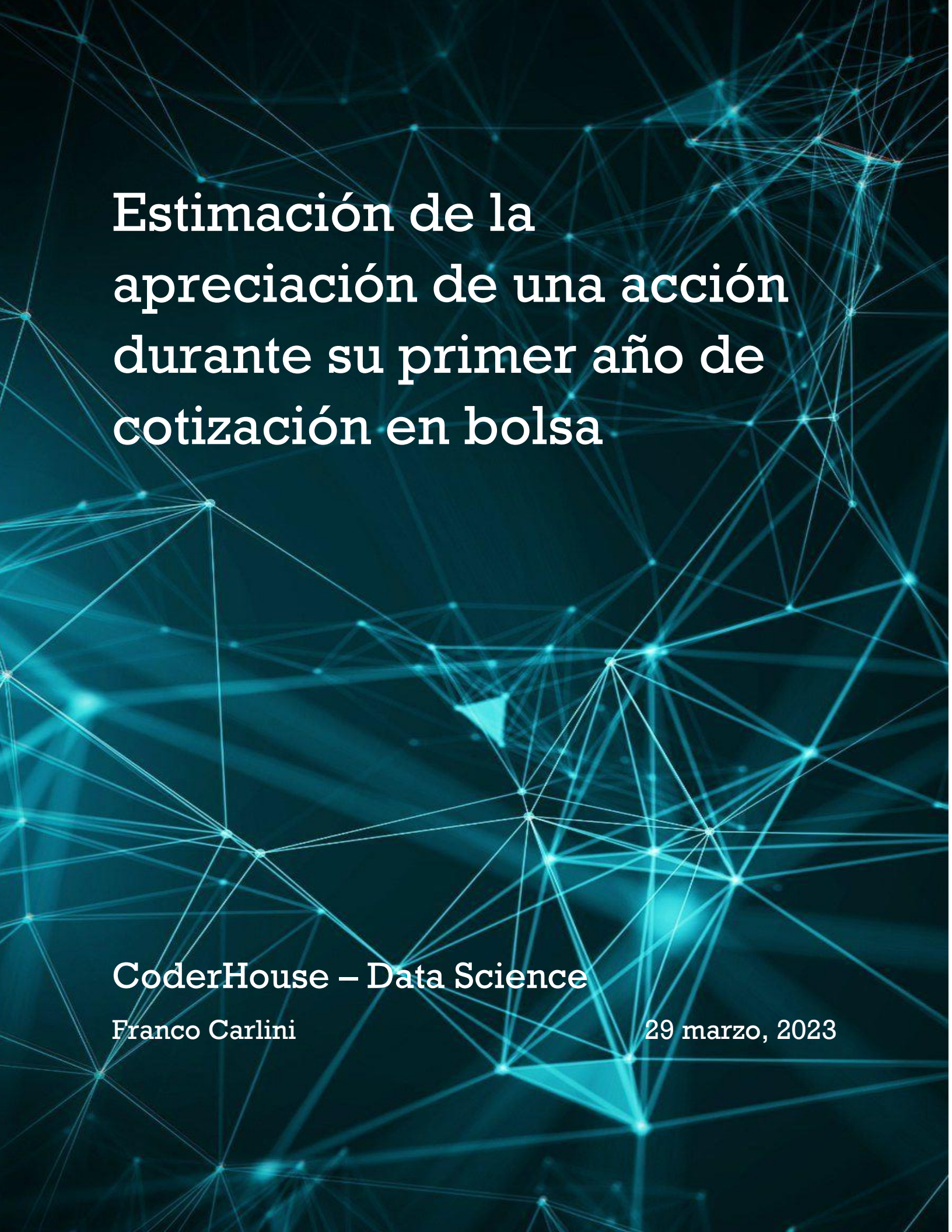


Estimación de la apreciación de una acción durante su primer año de cotización en bolsa

CoderHouse – Data Science

Franco Carlini

29 marzo, 2023

Tabla de Contenido

1. Motivación y Público Objetivo de la Investigación

Conforme pasan los años, conseguir retornos en inversiones en acciones que superen los rendimientos de sus competidores directos se hace más difícil. La teoría financiera indica que, acciones públicas que tengan una volatilidad parecida, variarán su precio de manera similar. Conseguir un retorno adicional entre productos financieros comparables es lo que se conoce como 'alpha'. En el contexto actual, conseguir este excedente es complejo debido a que compiten por él ejércitos de analistas que potencian sus habilidades mediante el uso de cada vez más rápidos e inteligentes algoritmos predictivos.

El interés predictivo y, por consiguiente, la mayor racionalidad de mercado, se concentra alrededor de los principales índices y compañías. Es alrededor de índices como el S&P por ejemplo, que conseguir un alpha superior al 0.5% es prácticamente imposible. Sin embargo, esto deja abierta la puerta de mercados menos conocidos donde haya menos competencia por parte de grandes entidades comercializadoras de acciones y por ende, precios menos racionales. Uno de ellos, y el que investigaremos en este documento, es el mercado de las empresas que recién salieron a cotizar tras una IPO.

El objetivo principal estará en determinar qué características de una empresa apenas finalizada su IPO son más importantes para predecir su desempeño en el mercado de valores. Esta investigación contempla más de 2000 empresas que salieron a cotizar entre 1996 y 2018, y cuyos rendimientos se analizan sobre su primer año de cotización.

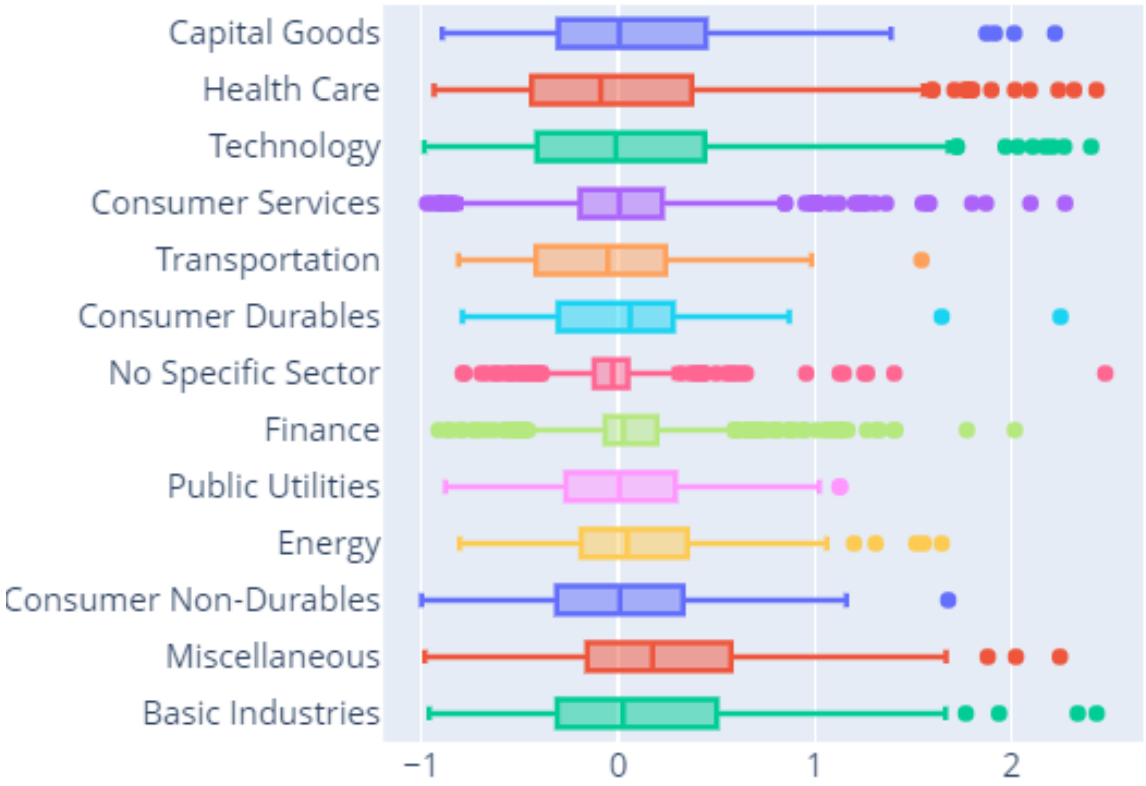
El impacto de esta investigación está dirigido a personas con acceso a vehículos de inversión complejos, pero que no cuentan con el capital suficiente o los contactos para acceder a IPO's y fondos sofisticados de manejo de capitales. En otras palabras, personas que han generado una buena bolsa de ahorros, pero no la suficiente para ser considerados influyentes en el mercado.

2. Hipótesis

1. Market Trend vs Fluctuacion del Precio

1.1. ¿Qué sectores lo hacen mejor en los años alcistas?

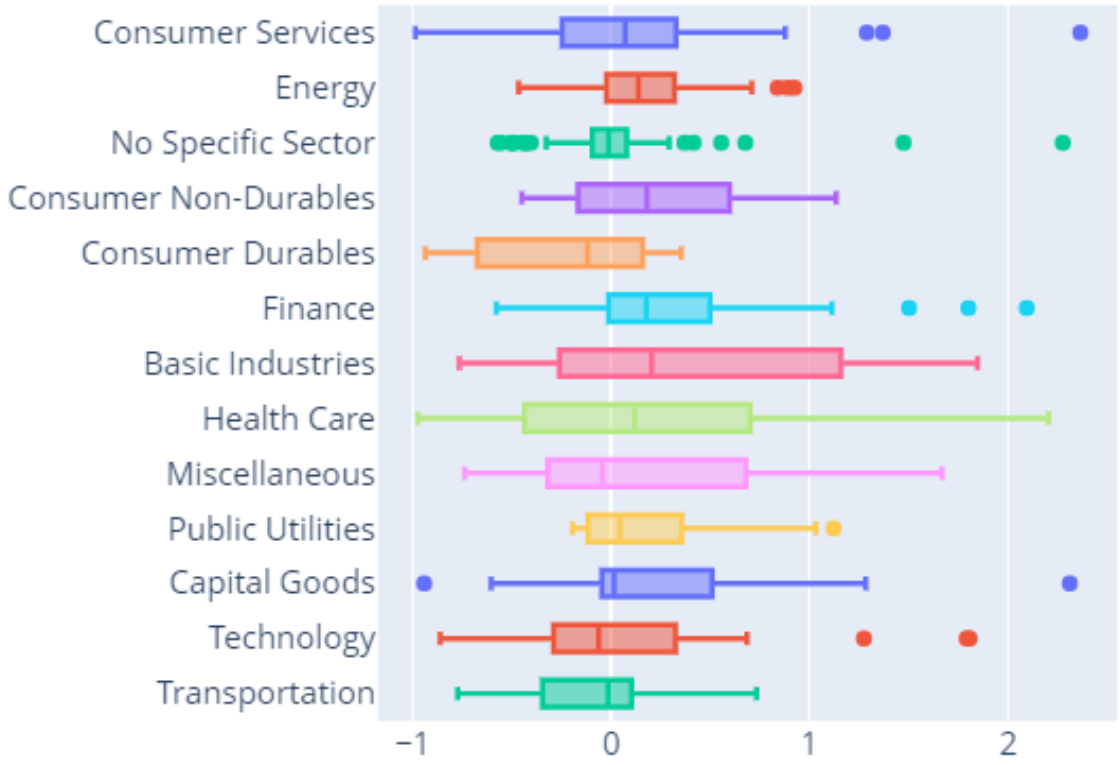
En años alcistas los resultados de los sectores suelen ser más centrados en torno a 0%. Sin embrago igual las medianas más altas se encuentran en 2 sectores específicamente: Misceláneos (22%) y Consumo Básico (o “Consumer Durables” con 6%).



1.2. ¿Qué sectores lo hacen mejor en los años bajistas?

En años bajistas importa bastante más el sector, porque hay sectores con muy buenos rendimientos y otros con muy malos. Del gráfico podemos concluir que esta vez 5 de los 13 sectores tienen medianas negativas cuando salen a cotizar en años negativos para el S&P , pero esta vez, son 5 sectores que tienen medianas mayores a 10% de rendimiento. Estos sectores son Basic Industries, Consumer Non-Durables, Finance, Health Care y Energy. Sorprende la mayor variabilidad en rendimiento de los sectores. Por un lado a más sectores les va mal, pero por otro, a los sectores que les va bien, les suele ir muy bien.

Además, de acuerdo con la teoría económica, los sectores cuyo consumo no puede ser diferido son a los que les suele ir mejor en años negativos para la economía. Esta vez la teoría y la práctica se encuentran poniendo a sectores como "Consumer Non-Durables" o "Basic Industries" a la cabeza de la lista con rendimientos muy altos. Además, el sector "Healthcare" lo suele hacer muy bien en estos entornos y la tabla lo demuestra con un crecimiento superior al 10%.



1.3. ¿Cuál es la proporción de empresas que salen a cotizar en años bajistas vs alcistas?

Para poder ponderar lo hallado anteriormente, es necesario saber qué escenario se repite con mayor frecuencia. Esto con fin de darle mayor énfasis a uno por sobre el otro. En teoría, deberían haber de 3 a 4 veces más años alcistas que bajistas, por lo que es más probable que muchas más empresas de nuestro dataset hayan salido a cotizar en años alcistas.

1.4. Las empresas de los diferentes sectores se distribuirán de forma normal en cuanto a crecimiento de la valorización se refiere.

La normalidad es un fenómeno que se observa recurrentemente en mediciones dentro de la naturaleza y del comportamiento humano. Debido a esto, existen diversas pruebas estadísticas que han sido diseñadas para ser usadas en poblaciones "normales". La ventaja de que una población sea normal, es que estas pruebas (conocidas como paramétricas), son mejores que las no paramétricas en cuanto a precisión se refiere.

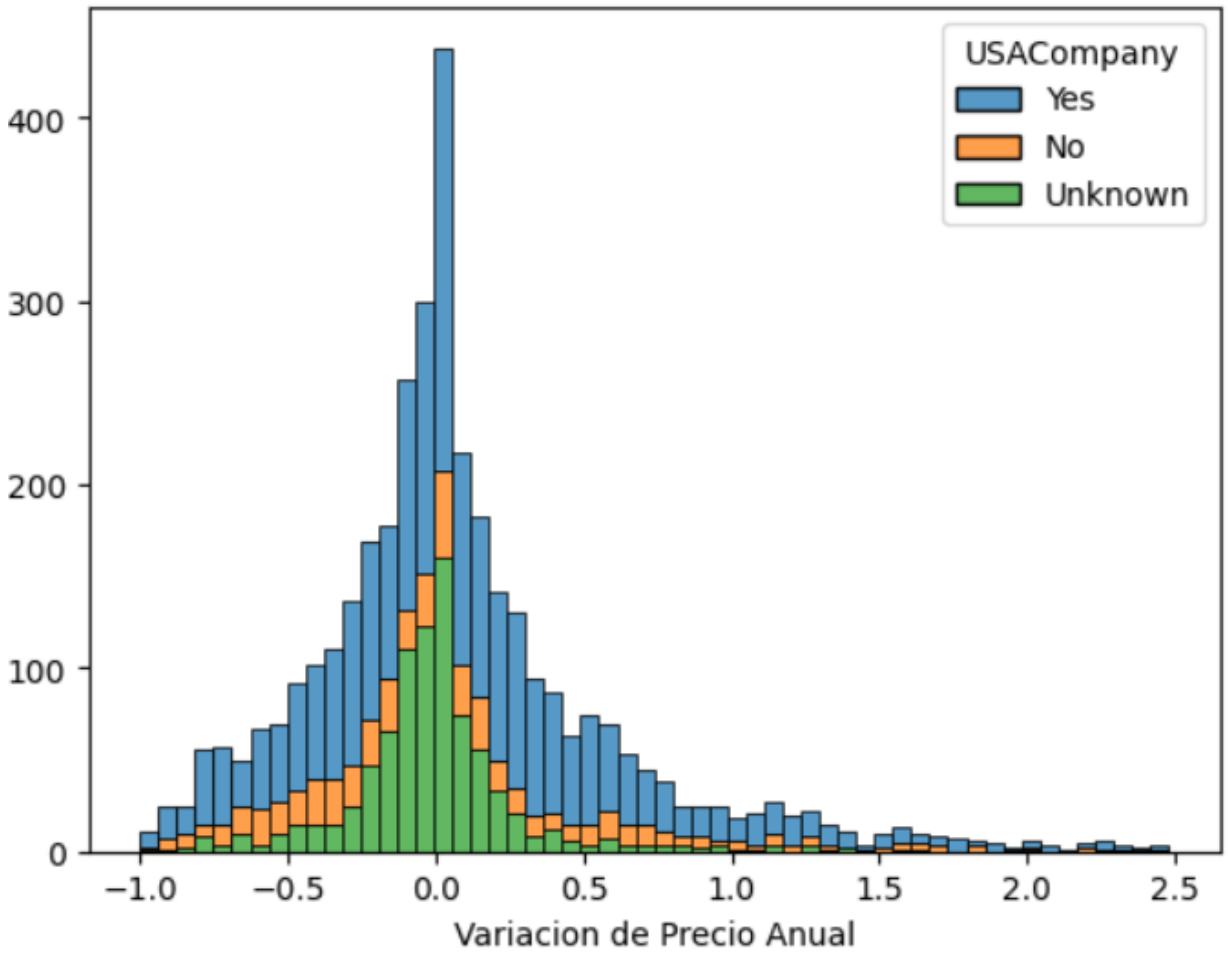
2.Ubicacion Geografica

2.1 ¿Las empresas de cuáles países se desempeñan mejor por sector?

Las empresas se concentran principalmente en EE.UU. (61%) y solo 15% se ubican en el extranjero.

Las empresas de EE.UU. tienen una distribución de datos más positiva, es decir tienen mejor desempeño en promedio. La media y mediana de los grupos geográficos son:

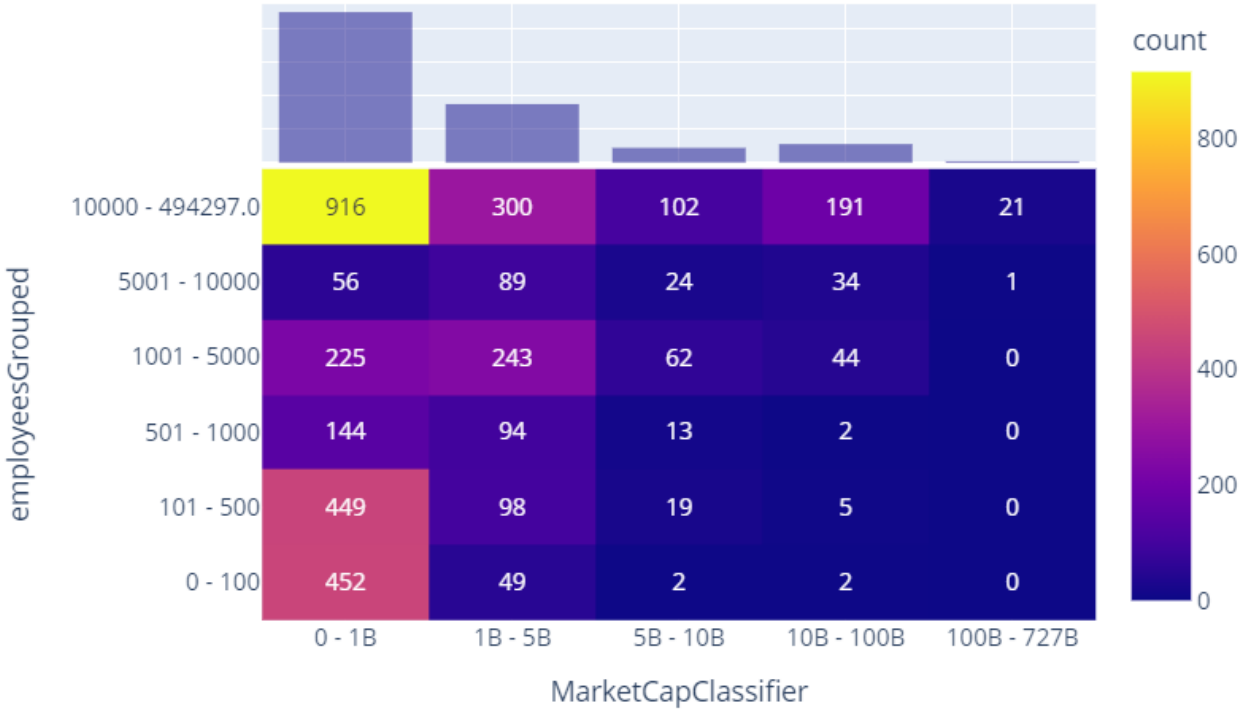
- EE.UU: 11% y 3%
- Resto del Mundo: 6% y -1%
- Desconocido: 0% y -1%



3. Empleados

3.1 ¿La cantidad de empleados se relaciona con el Market Cap?

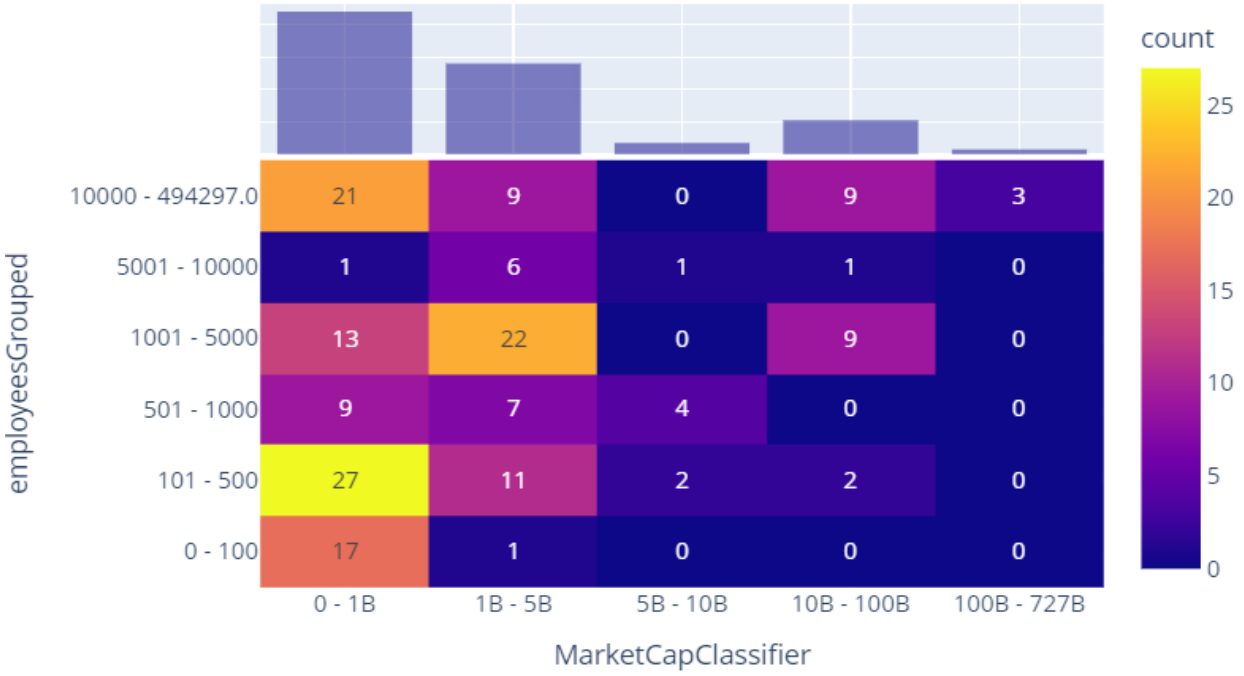
Distribución de empleados en todos los sectores



Las empresas de todos los sectores suelen seguir la tendencia general de mayor escala igual a mayor número de empleados. Sin embargo, los gráficos muestras que efectivamente sectores como la tecnología y el de Energía, no la siguen tan de cerca. Este hallazgo es interesante porque revaloriza escalar sin mayor inversión adicional en mano de obra. Considerando que es de los mayores costos de muchas empresas, significa que al crecer, compañías de estos sectores rápidamente expanden sus márgenes.

Sector Tecnología

Sector Energía



4. Longevidad de la Empresa

4.1 Una empresa antigua tenderá a hacerlo mejor en su primer año de cotización en mercados públicos?

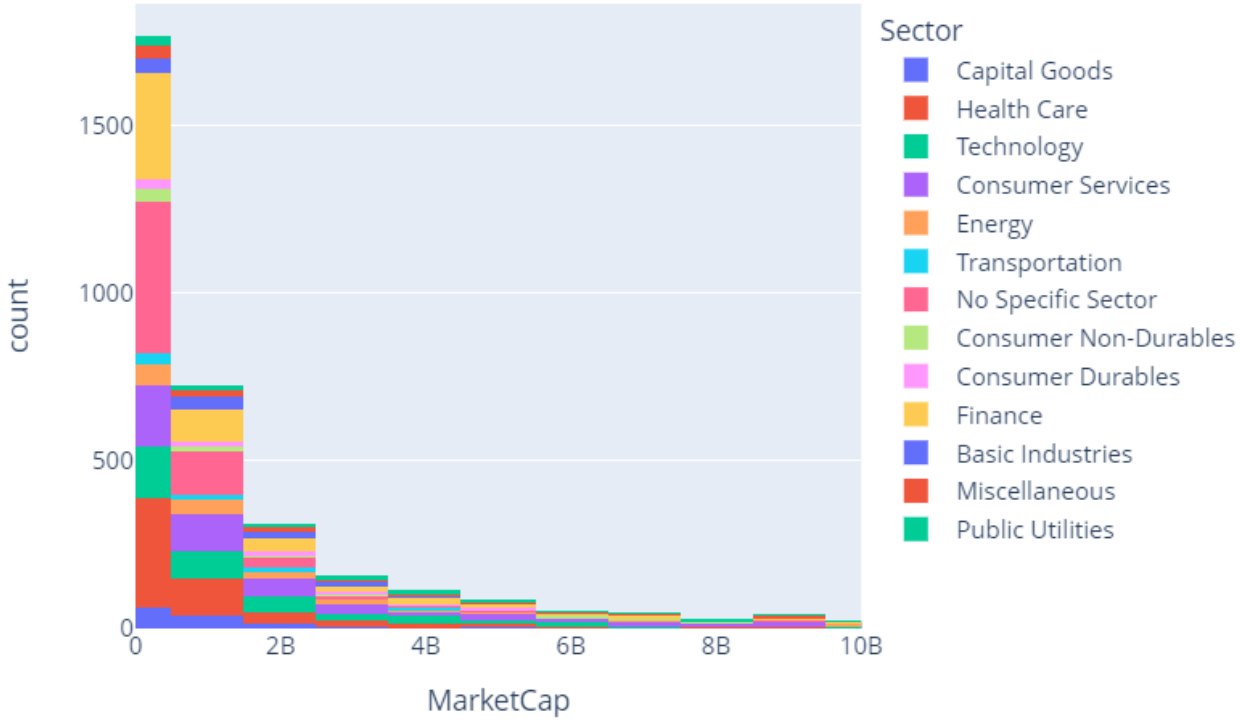
Lógicamente una empresa que viene siendo gestionada por más años antes de salir a cotizar en bolsa debería estar mejor organizada y por ende verse como una opción más sólida para invertir que una empresa que no a sobrevivido a la "prueba del tiempo". Un comportamiento relacionado a la solidez no necesariamente es hacerlo mejor en términos de incremento de valorización, sino, qué tanta variabilidad hay entre empresas nuevas vs las experimentadas.

Según el estudio, una empresa más antigua no lo hará necesariamente mejor en bolsa en su primer año de cotización, pero sí será más predecible el rango sobre el cual fluctuará el incremento en su valoración según el gráfico. Además, viendo el tamaño de las burbujas, parece que las empresas con mayor Market Cap suelen tener también un rango más acotado. Con estos dos hallazgos podemos suponer que si lo que se busca son rendimientos más predecibles pero menores, empresas grandes y con más edad son la mejor opción cuando salen a cotizar.

Evaluando los gráficos, tiene sentido que una empresa nueva rinda más con mucho mayor riesgo debido a que tiene un mayor potencial de crecimiento. El tema de sobrevivir la "prueba del tiempo", parece ser relevante en cuanto a un rango más acotado de incremento de valorización.

5. Volatilidad

5.1 ¿Cómo se distribuye la capitalización de mercado de las IPO's?



Como se ve en el gráfico de distribución, los datos se encuentran sesgados a la derecha, es decir, apilados cerca al eje 'Y'. Esto tiene sentido ya que es muy difícil alcanzar una capitalización de mercado muy alta y a menudo unas pocas empresas se suelen quedar con desproporcionalmente más mercado ("the winner takes all").

5.2 ¿La capitalización de mercado está inversamente relacionada con la variación de la cotización anual y la volatilidad diaria?

Tiene sentido esperar que la relación entre Market Cap y la variación en la cotización anual sea inversa. Esto debido a que a mayor valor, más masa monetaria (ingresos o salidas de capital) se necesitarán para mover el precio de una acción. Sin embargo, es preciso tener en cuenta que en última instancia la liquidez es el factor determinante. Por ejemplo, una empresa muy grande puede fluctuar mucho en bolsa si muy poquitos de sus inversores desean vender. Los resultados se muestran a continuación:

Utilizando el MarketCap Original

	MarketCap	Year_Price_Variation	Stan Dev.
MarketCap	1.000000	0.033947	-0.004225
Year_Price_Variation	0.033947	1.000000	0.017351
Stan Dev.	-0.004225	0.017351	1.000000

Vemos que el Market Cap no está correlacionado ni con la variación anual, ni con la volatilidad diaria. A continación, veremos si el Market Cap visto por categorías (MC discreto) sí tiene alguna correlación al menos superior a 0.3.

Utilizando la variable modificada MC_discrete_market_classifier

	MC_discrete_classifier	Year_Price_Variation	Stan Dev.
MC_discrete_classifier	1.000000	0.053283	-0.014762
Year_Price_Variation	0.053283	1.000000	0.017351
Stan Dev.	-0.014762	0.017351	1.000000

Como se ve en la tabla, si bien hay pequeñas variaciones, ninguna correlación puede ser ni siquiera considerada como 'débil'. La correlación entre la variación de precio anual y la desviación estándar diaria será evaluada más adelante.

5.3 ¿Mayor volatilidad tiende a tener mayores retornos? Evaluar correlaciones por sector.

Debido a que ningún sector obtuvo una correlación mayor a 0.38, podemos decir que otra vez vemos que no existe ninguna correlación significativa entre la variación anual y la volatilidad. Esta hipótesis demuestra que ningún sector posee esta correlación y por lo tanto, para este dataset, esta hipótesis es falsa.

5.4 ¿Hay relación entre el volumen diario medio como porcentaje del Market Cap de una acción y su volatilidad diaria?

	Year_Price_Variation	proportion_daily_vol_of_MC
Year_Price_Variation	1.000000	0.047397
proportion_daily_vol_of_MC	0.047397	1.000000

En teoría, a mayor número de acciones comercializadas debería haber menor volatilidad en los precios de la misma. Sin embargo, la proporción del valor de las acciones vendidas en relación con el Market Cap de la compañía no está correlacionada con la variación anual del precio de la acción.

6. Ratios financieros: P/E to growth

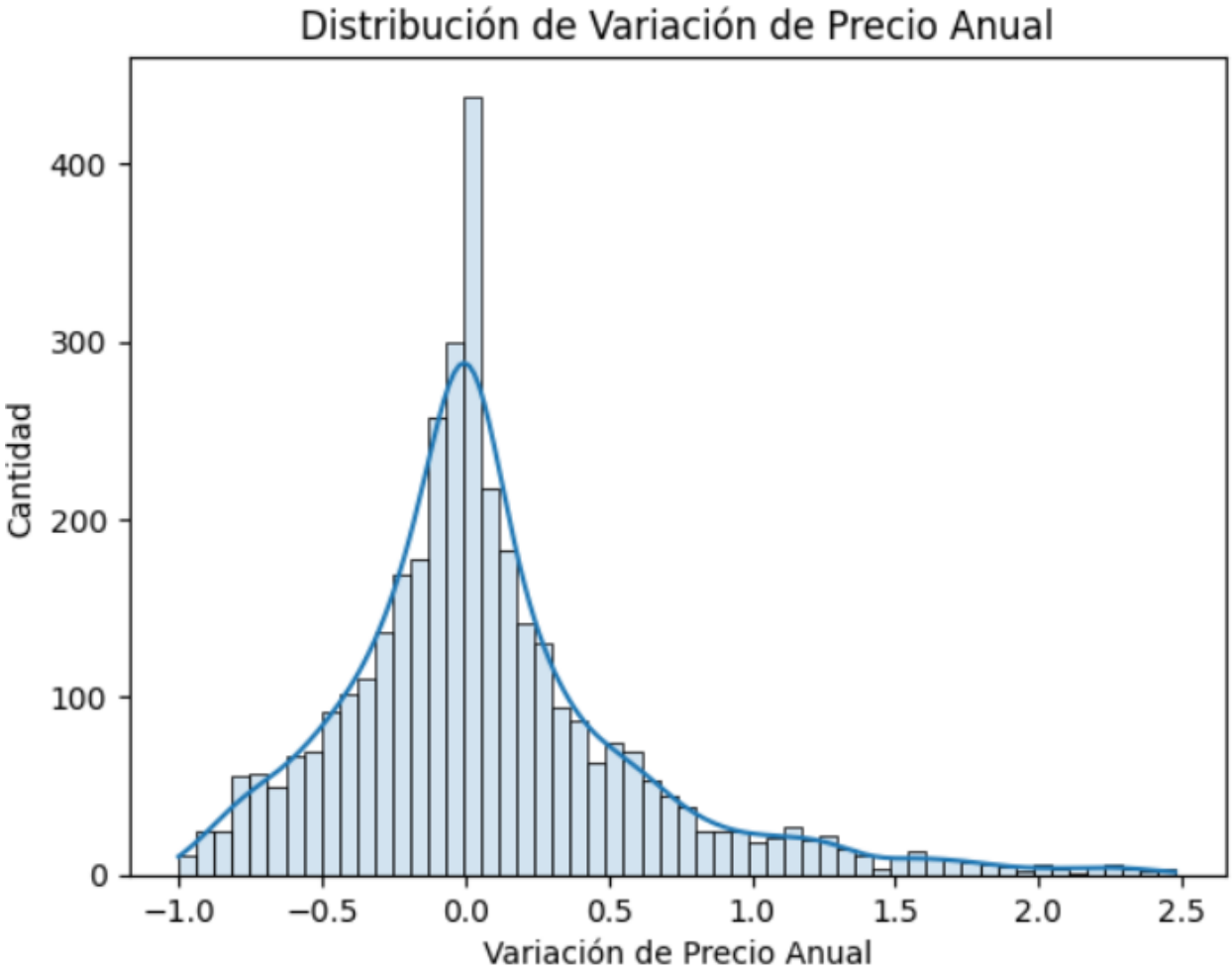
6.1 ¿Cuál es la relación entre el P/E del año anterior a cotizar con la cotización en el primer año en bolsa?

El P/E es un indicador del precio de una acción relativo al número de veces de esta sobre los EPS (ingresos netos por acción). Ya que indica el "número de veces" forma parte de los indicadores conocidos como "múltiplos". Generamente un P/E por debajo de la media indica que la empresa está subvaluada y por ende tiene más probabilidades de aumentar su cotización en el futuro. Deberíamos esperar una correlación negativa entre el P/E del año antes de salir a cotizar con la variación del precio en bolsa.

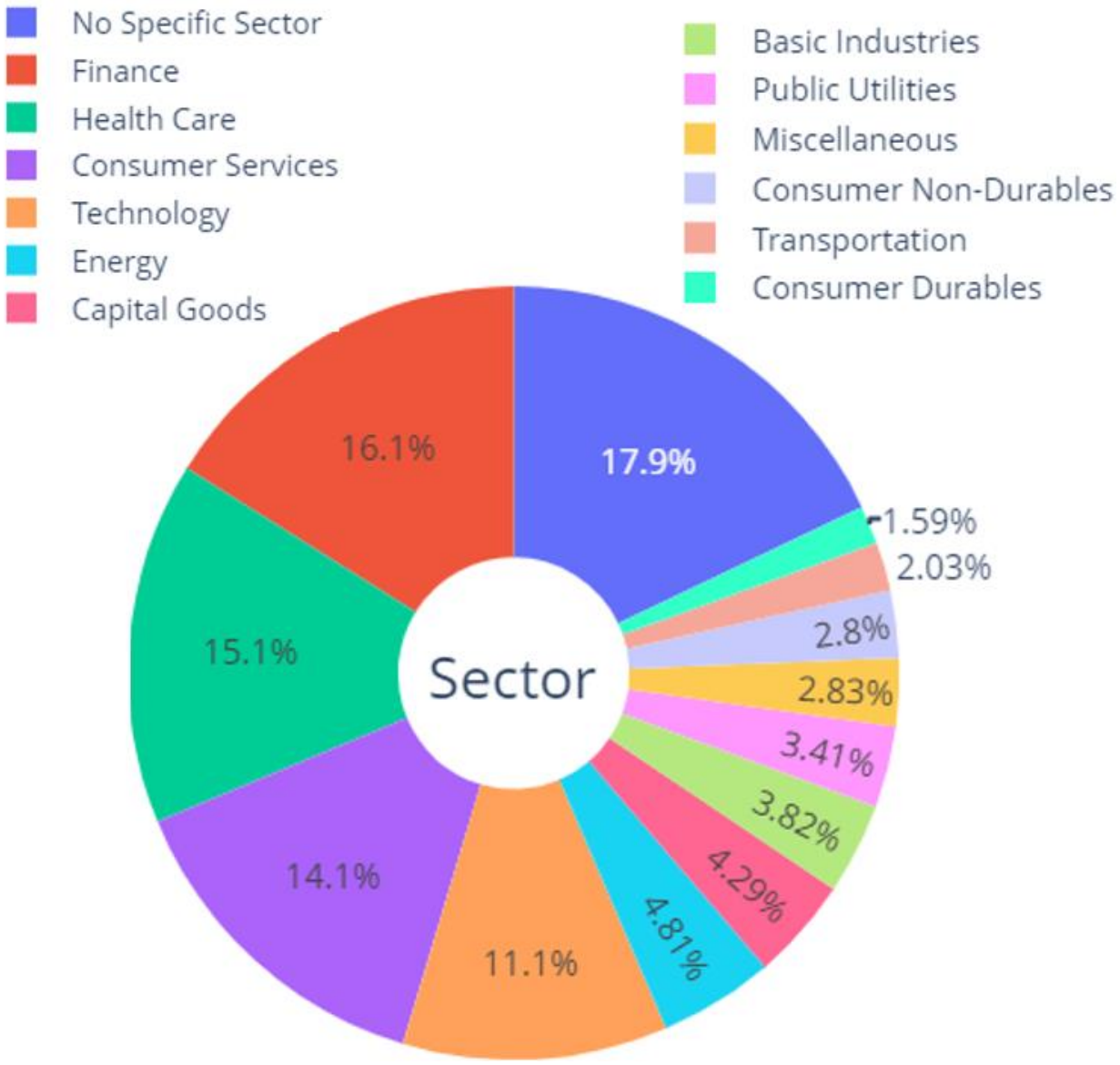
En Health Care, Technology, Energy y Miscellaneous parece comprobarse el efecto inverso del P/E sobre la variación del precio (a menor P/E, mayor potencial de revalorización). Sin embargo los resultados en general son decepcionantes en cuanto a esta hipótesis. Parece que en el caso de empresas que recién salen a cotizar, el P/E del año previo a salir a bolsa no parece ser muy relevante, puede deberse a que no hay un historial público como para determinar si la empresa está cara o barata respecto a su histórico y solo se le puede contrastar con el sector. Otra explicación es que la euforia de la salida a bolsa minimice el efecto de una valoración razonble basada en el múltiplo P/E. Finalmente, quizás este indicador sí tenga un peso significativo ya que hay sectores donde parece influir, pero no sea algo muy importante

3. Análisis exploratorio de datos (EDA)

Este estudio considerae lo visto previamente en el análisis de las hipótesis como parte del EDA. Como adicional se observa que el dataset depurado cuenta con 1902 registros y 27 variables que se distribuyen como se ve en el siguiente histograma. La media aritmética es de 7.5%, la mediana de 0.7% y la moda se ubica en el intervalo entre 0 y 0.625%.



Las empresas se distribuyen por sector, que como se verá en el resto del informe, tienen un peso importante sobre la proyección del rendimiento en bolsa de una compañía. Los sectores dominantes según el gráfico son Finanzas, Salud Servicios y Tecnología, entre los 4 concentran el 56.3% de las empresas. Además, es preciso mencionar que 18% de compañías no tienen un sector asignado.


No Specific Sector
Finance
Health Care
Consumer Services
Technology
Energy
Capital Goods
Basic Industries
Public Utilities
Miscellaneous
Consumer Non-Durables
Transportation
Consumer Durables
Sector
17.9%
16.1%
15.1%
14.1%
11.1%
4.81%
4.29%
3.82%
3.41%
2.83%
2.8%
2.03%
1.59%

4. Ingeniería de Atributos

Limpiamos la base de datos para realizar la reduccion de dimensionalidad

Hay variables que debemos volver numéricas para poder incluirlas en la matriz de correlación. Estas son:

1. netIncome
2. Revenue
3. employeesGrouped
4. CEOGender
5. USACompany
6. FiscalMonth

Las que son puramente categóricas salvo "Sector" cuentan con muy pocos datos en cada una de sus categorias como para hacer un modelo preciso. Por lo cual seran eliminadas

1. Industry
2. City
3. stateCountry

Hay variables categóricas ordinales que es importante evaluar. Estas parten de variables cuantitativas y la razón de su inclusión en esta parte del estudio es que permiten evaluar una misma variable cuyo efecto puede ser no linear, de manera lineal.

1. MC_discrete_classifier -> Se descarta MarketCap porque sería redundante en el modelo (el tamaño de la empresa sería representado por dos variables, minimizando el efecto de ambas).

2. employeesGrouped (primero debemos volverla discreta ordinal) -> Se descarta employees porque sería redundante en el modelo (el numero de empleados de la empresa sería representado por dos variables, minimizando el efecto de ambas).

Finalmente, variables a descartar (ademas de las categoricas mencionadas):

1. dayOfWeek - irrelevante: el día de la semana de la IPO no es un fundamento sólido para cotizar una variable.

2. Summary Quote - irrelevante: el enlace web nada tiene que hacer con la performance de una variable.

3. FiscalDateEnd - irrelevante/obsoleta: La fecha exacta del fin del año fiscal no es relevante, sobre todo si ya hay otra variable que indica el mes del fin del año fiscal (que quizás si sea relevante para la planificación de la empresa).

4. YearFounded - obsoleta: YearsBeforeIpo la reemplaza en relevancia.

5. Year - obsoleta: YearsBeforeIpo la reemplaza en relevancia.

6. DaysBetterThanSP: No es considerada predictiva puesto que se evalua en base al rendimiento anual de la variable objetivo.

7. MarketYearTrend: No es considerada predictiva puesto que se evalua en base al rendimiento anual del mercado.

8. Median, Mean y Stan. Dev. : No son consideradas predictivas puesto que se evaluan en base al rendimiento diario de la variable objetivo.

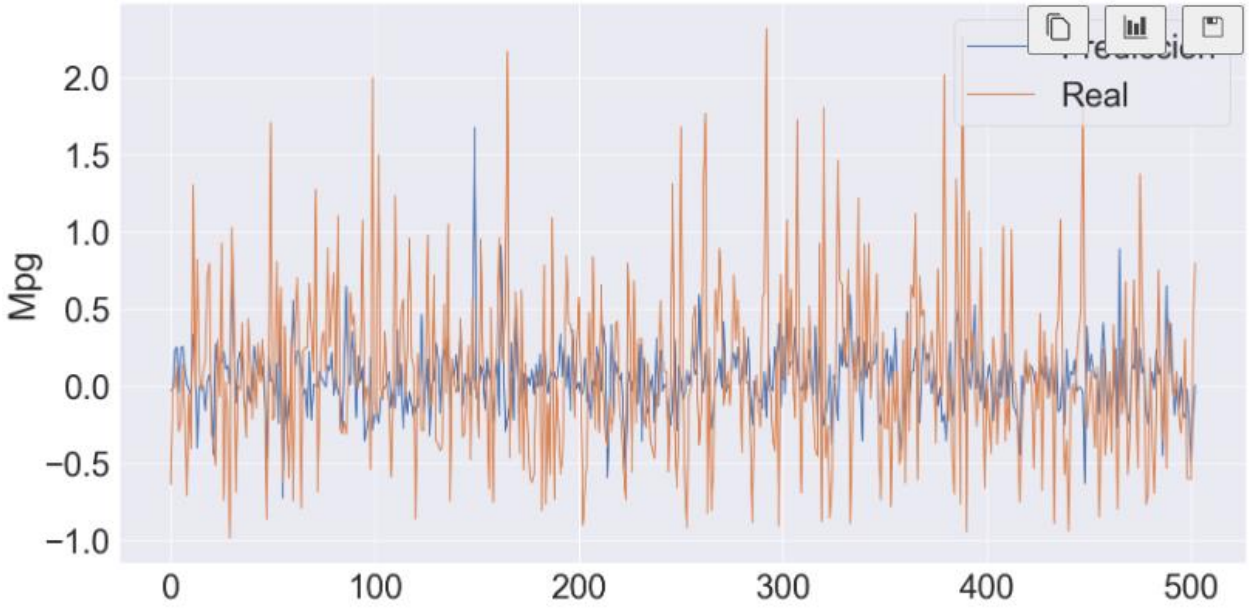
9. Retiro de la variable 'Safe' (presumo que hace referencia al termino financiero "Simple Agreement for Future Equity") por no tener claridad en la definicion por parte de la fuente del dataset (Kaggle).

Además, generamos variables “dummy” para la variable discreta “Sector”.

5. Entrenamiento y testeo

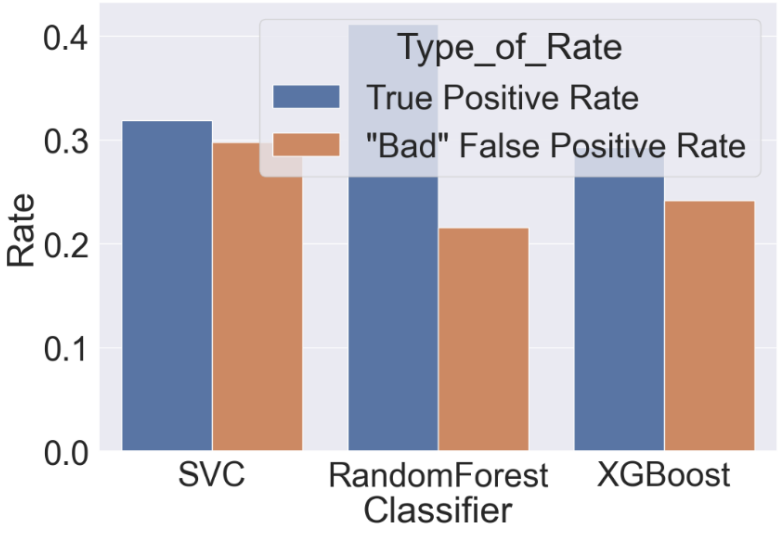
Debido a que ninguna regresión se ajusta bien al Dataset (Ridge, RidgeCV, Random Forest, XGBoost y SVR), incluso la mejor (SVR) tiene un error porcentual (MAPE) muy elevado (207%). Por este motivo se deberá testear con algoritmos de clasificación, para lo cual primero se crearán categorías.

El error se debe a que los valores reales son muy extremos (como se ve en la gráfica) y varían mucho en cuanto a las variables predictoras.

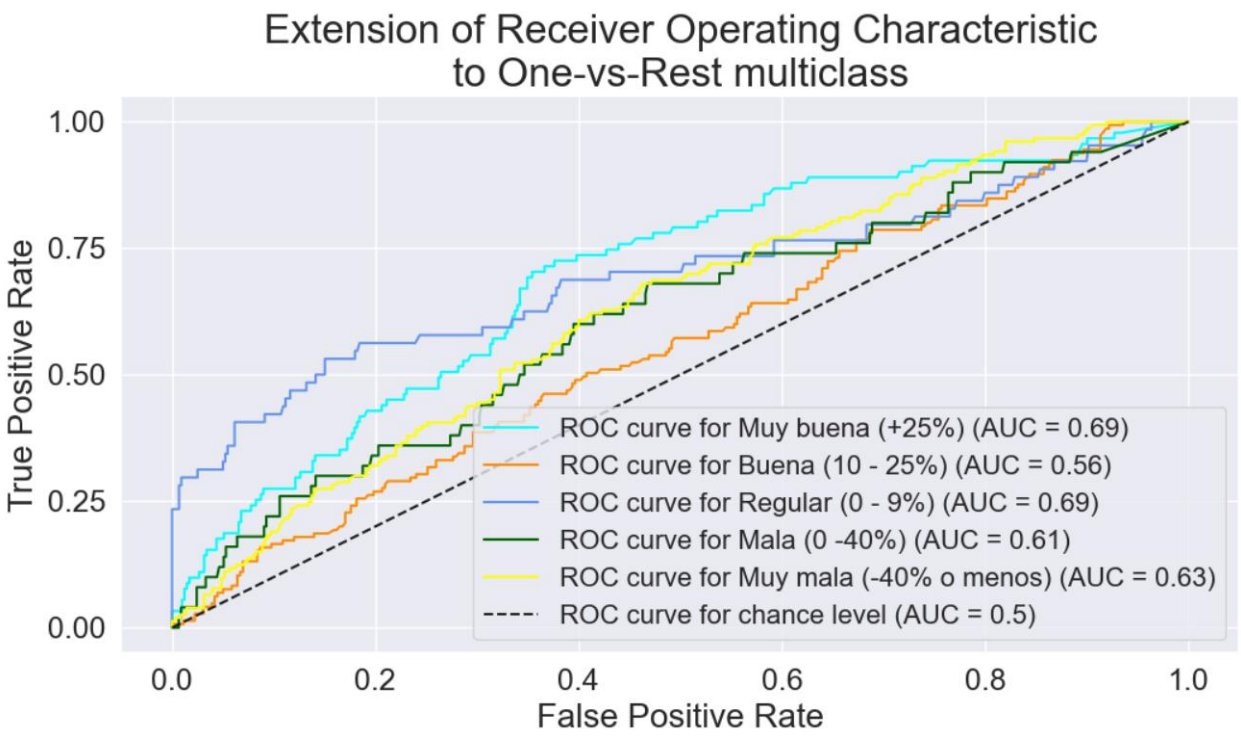


Para la clasificación probamos 4 puntos de corte para tener 5 categorías para la performance de las compañías:

1. Muy Buena (+25%)
2. Buena (10 a 25%)
3. Regular (0 a 9%)
4. Mala (-40 a 0%)
5. Muy mala (-100 a -40%)



El Mejor modelo es el RandomForest, por lo que se graficó su curva ROC múltiple a continuación.



Debido a que la decisión del público objetivo (inversores) es binaria (invertir o no invertir), se determino el mejor punto de corte para evaluar esta decisión. Sin embrago, recordemos que la clasificación binaria (simple) puede traer el peligro de que se considera igual una compañía que crece 9% a una que quiebra (-100%). Por lo que trazaremos los valores promedio de los True Positives y los False Positives para determinar si el valor esperado es rentable.

corte	Punto de Corte	TP Mean	FP Mean	Precision	Score
0.00	Corte en 0%	0.451117	-0.302203	0.604938	0.153509
0.10	Corte en 10%	0.553301	-0.246863	0.542553	0.187269
0.25	Corte en 25%	0.619160	-0.099845	0.405797	0.191925

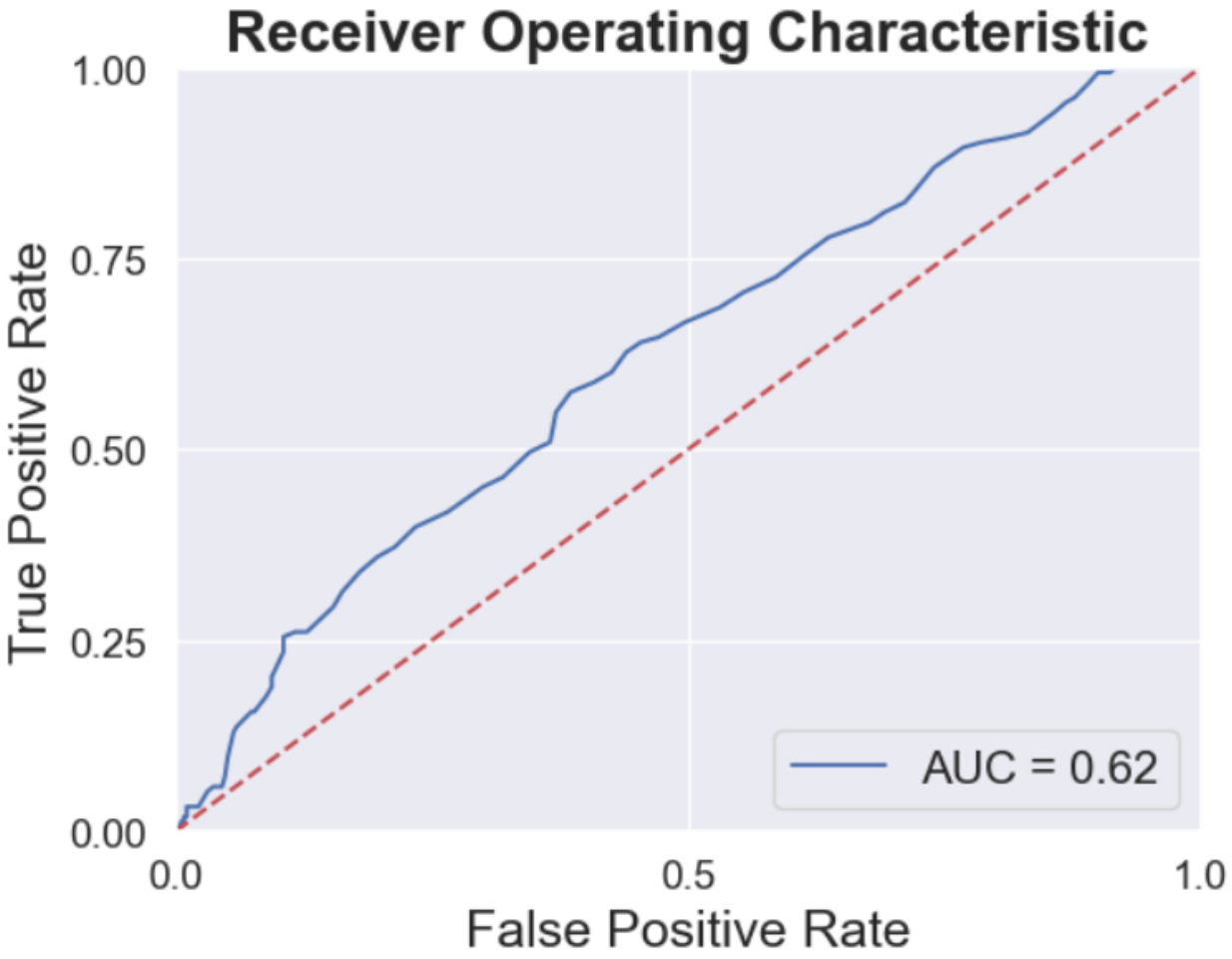
*Score: es el rendimiento promedio de un portafolio que contiene todas las entradas de prueba (test_X) donde se consideran las magnitudes de los rendimientos.

De los experimentos realizados, se eligen dos puntos de corte frecuentemente, 10% y 25%. En este informe usaremos 10%.

El modelo predice poco más de la mitad (53%) de las veces acertadamente si una acción incrementará su valor más de un 10%.

Según la matriz de confusión, si el modelo predice que será positivo, hay 47% de posibilidades (86/183) que se trate de un falso positivo. En el caso del score "False Positive Rate", la probabilidad de un falso positivo es 44.6% (cuando se hace cross-validation). Este error (tipo 1, falso positivo) es más riesgoso que el tipo 2 (falso negativo) para este tipo de inversión, ya que una mala inversión puede equivaler a entre 9 y 10 veces buenas inversiones (debido a que puede ser entre -100% y + 9 %).

El modelo es bueno porque otorga un rendimiento promedio aproximado de 20% anual. Esto es muy superior al retorno de ETF de mediano/gran rendimiento como el del S&P500 (9%).



Además, según el feature importance classifier, las dos variables más importantes son LastSale y NetIncome.Es curioso porque estas no son relativas a la capitalización de mercado de la compañía, sino que son valores absolutos. Es decir, parece favorecer empresas con mayor capitalización de mercado. Recordemos que según el cuadro de correlaciones, LastSale tiene una correlación de 0.5 (mediana correlación) con MC_discrete_classifier, pero no tiene ninguna correlación con NetIncome.

	feature_names	feat_imp	std
3	LastSale	0.045	0.008
0	netIncome	0.021	0.006
2	YearsBeforeIpo	0.012	0.011
1	MC_discrete_classifier	0.008	0.002
16	Miscellaneous	0.006	0.003
14	Finance	0.005	0.004
5	USACompany	0.003	0.003
12	Consumer Services	0.003	0.004
15	Health Care	0.002	0.002
10	Consumer Durables	0.002	0.001
4	Revenue	0.001	0.009
17	No Specific Sector	0.001	0.005
20	Transportation	0.001	0.002
11	Consumer Non-Durables	0.000	0.003
8	Basic Industries	-0.000	0.002
6	FiscalMonth	-0.000	0.007
13	Energy	-0.001	0.003
7	employeesGrouped_ordinal	-0.001	0.006
18	Public Utilities	-0.002	0.003
19	Technology	-0.004	0.003
9	Capital Goods	-0.005	0.003

6. Optimización

Se probó con GridSearchCV y RandomizedSearchCV para los dos mejores modelos binarios (RandomForest y XGBoost)

RandomForest:

La selección de parámetros es diferente:

1. n_estimators: El HalvingRandomSearch maximiza el número de árboles mientras que el RandomSearch lo deja en 5 (el mínimo).

2. criterion: Ambos prefieren el log_loss.

3. max_depth:La profundidad del Halving es 3 (menor), mientras que la del RandomSearch es 5 (intermedia).

4. bootstrap: Ambos están de acuerdo con hacer bootstrapping (True)

5. min_samples_split: El Halving prefiere un mínimo de 1 mientras que el RandomSearch de 2.

Mejor accuracy por parte del HalvingRandomSearch (69.3% vs 68.9%) . Sin embargo, demoró más. Extraño que siendo el CV Score el Accuracy, su valor no coincida en ningún caso con el de la función accuracy.

7. Selección de modelo

El modelo elegido es sin duda el RandomForest Classifier. Seguido de cerca por el XGBoost. Algunas recomendaciones para inversores son las siguientes:

- No confiarse creyendo que la gran mayoría de acciones nuevas producen grandes ganancias y hacer un buen *stock picking*.
- Darle mayor peso a los hallazgos de los sectores Financier, Salud, Servicios y Tecnología.
- Priorizar la compra de acciones de EE.UU.
- Invertir en años bajistas en Consumo Estable, Industrias Básicas, Energía, Financiero y Salud. Invertir fuera de estos sector puede inciter grandes pérdidas.
- En años alcistas darle prioridad a Consumo Discrecional y a Misceláneo.
- Invertir en empresas que escalen fácilment (no tengan la necesidad de adquirir mucho personal para crecer). De preferencia priorizar los sectores Variado, Servicios y Tecnología.