ITESO - Maestría en Sistemas Computacionales

Asesor: J. Guadalupe Olascuaga Cabrera

Co asesor: Luis Fernando Gutiérrez Preciado

Alumno: Mawrer Amed Ramirez Martinez

Reporte de Avance de Trabajo de Obtención de Grado

Asesoría: 4 Junio de 2019 – IDI4

**Completado:**

* **~~Curso de Cursera. Python for Financial Analisis. (análisis de Features y porque ciertos features).~~** [~~https://www.coursera.org/learn/python-statistics-financial-analysis/home/welcome~~](https://www.coursera.org/learn/python-statistics-financial-analysis/home/welcome)

**Avance:**

Finally Evaluate the performance of the model. Stadistics Standards: Sharpe Ratio and Maximum Drawdown.

Definicion:

Manejo de Jupiter Notebooks: guias y best practices: <https://www.datacamp.com/community/tutorials/tutorial-jupyter-notebook>

**Packages for python:**

Pandas. (Dataframe and series). Handling missing value, Computing Pariwise correlation.

Numpy. Generate random numbers.

Matplotlib.

Statsmodels (regressions and time series analysis)

fb = pd.**from\_csv**('../data/facebook.csv',index\_col=0)

Basics attributes of Dataframe:

.head()

.index

.index[0]

.index[-1]

.shape

.tail()

.describe()

Slicing DataFrame

Selection by Label:

.loc

Selection by position

.iloc

ms.loc['2016-01-01':'2016-12-31','Close'].plot()

Generation of Variables:

ms[‘Close’].**shift**(-1)

ms[‘**PriceDiff**]= ms [‘price1’] – ms[‘Close’]

**Dailyreturn** = priceDiff / Close

**Direction** = PriceDiff = 1 or -1. (Compra o venta)

**Ms[‘Direction’] = [1 if ms.loc[ei, ‘PriceDiff] > 0 else -1 for ei in fb.index]**

Moving Average = using .rolling()

MS[‘MA40’] = MS[‘Close’].rolling(40).mean()

Average3 = Shift(2)

Semana2:

Random Variables, Distribution in Random variables.

Models of Stock Return.

**Daily Return of Stock Price**

Frecuency and distribution.

.value\_counts()

.sort\_index()

.mean and Variance to describe the distribution of probability.

A screenshot of a social media post

Description automatically generated

**Continuous Random Vairables.**

Normal Random Viariable.

Scipy.stats import norm

Density = pd.DataFrame()

Density[‘x’] = np.arange(-4,4,0.001)

Density[‘pdf’] =norm.pds(density[‘x’],0,1) #get PDF

Density[‘cdf’] =norm.cdf(density[‘x’],0,1) # get CDF

***PDF*** *Probability Density Function*

***CFD*** *Cumulative Distribution Function*

A screenshot of a cell phone

Description automatically generated

Value at Risk.

A screenshot of a cell phone

Description automatically generated

*“Is it safe to use normal description to model stock return?*

*Two famous professors in the field asset pricing, Fama and*

*French responds in this way.*

*Distribution of a daily and monthly stock return,*

*are rather symmetric about their means, but the tails are fatter.*

*Which means there are more outliers that would be*

*expected with normal distributions.*

*It means that, if tail returns negative,*

*as well as positive, may occur more often than we expect.*

*I” [1]*

# **Leer: Fama–French three-factor model**

"The Cross-Section of Expected Stock Returns"

<https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=1455>

<http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/consul_rel.html>

<https://books.google.com.mx/books?id=0Fs8DwAAQBAJ&dq=+Common+Risk+Factors+in+the+Returns+on+Stocks+and+Bonds&lr=>

Bibliografia:

[1]

“Kenneth R. French - Data Library,” *Dartmouth.edu*, 2019. [Online]. Available: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data\_library.html. [Accessed: 29-May-2020].

‌Quiz II:

Discrete Random Variable or Continuous Random Variable?

Relative Frecuency.

A screenshot of a cell phone

Description automatically generated

Sampling and Inference.

**1 Population and Sample**

**2 Variation of Sample**

**3 Confidence Interval**

**4 Hyposthesis Testing.**

Caracteristicas de la población.

Return of Investment

1 Population and Sample

All registers that share characteristics.

Random Sampling (with replacement or without replacement).

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Ddof=0 significa que N sera el tamaño total de la poblacion

A screenshot of a social media post

Description automatically generated

Muestra de la población:

Notice that when we compute a sample variance and

std, we need the ddof equal to 1.

It means the denominator of a sample variance has to

be n-1 instead of n, which is the sample size

ddof = denomitator of sample variance

**Degrees of Freedom.**

Number of values in calculation that are free to variate.

A screenshot of a cell phone

Description automatically generated

Variation of Sample:

***Incluso si una población no parece tener una distribución normal, si se hacen miles de muestreos en base a esa población y se grafican, entonces se notara como si existe una distribución normal. :o***

Confidence Interval

A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

Hypotesis Testing:

Null Hipotesis

Alternative Hypothesis:

T distribution

The *t-*distribution can be thought of as a cousin of the standard normal distribution — it looks similar in that it’s centered at zero and has a basic bell-shape, but it’s shorter and flatter around the center than the *Z-*distribution. Its standard deviation is proportionally larger compared to the *Z,* which is why you see the fatter tails on each side.

Z Distribution standard normal (or *Z-*distribution)

is the most common normal distribution, with a mean of 0 and standard deviation of 1

Comparing the standard normal (<i/>Z-) distribution to a generic <i>t-</i>distribution.”/>
<div class=

Standarization

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Modulo 4:

Focus on consistency of performance of modules, how to evaluate it correctly?

Test performance.

Association of multiple variables.

En estadistica se utiliza Covarianza.

**Coefficient of correlation. Nos permite encontrar esa relacion entre dos variables.**

df.cov()

df.corr()

Simple Linear Relation Model.

A screenshot of a cell phone

Description automatically generated

A close up of a logo

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

# scatter matrix plot

from pandas.tools.plotting import scatter\_matrix

sm = scatter\_matrix(housing, figsize=(10, 10))

Importancia de R2, ayuda a medir el performance de un modelo. Debo entender que mi TOG no esta enfocado en mejorar el performance de un modelo, sino en aplicar una estrategia de trading con ganancias claras usando uno o varios modelos.

Index SPY.

Identificar relación de timezones y cuando abren los mercados, Quizas ver volatilidad, identiciar cuando el spread es menor. Analisis dinámica del mercado durante horarios y días.

Prediccion, utiliza multiples predictors.

A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

Noise of the markets:

A screenshot of a cell phone

Description automatically generated

Con la correlacion podemos determinar los índices que tienen mas impacto con SPY:

A screenshot of a cell phone

Description automatically generated

Revisar P Value: (significacia del modelo).

A screenshot of a cell phone

Description automatically generated

IMPORTANTE DETERMINAR QUE EL MODELO NO ESTA SOBREENTRENADO, COMPARANDO R2, RMSE EN TRAIN Y TEST DATASETS.

A screenshot of a cell phone

Description automatically generated

A picture containing bird, knife

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**Trabajo para la siguiente sesión:**

* Entender bien el modelo LSTM (porque 51, porque 25).
* TENER EL INTERVALO que voy a utilizar
  + Diferencia entre gradiente deciendiente y gradiente estocástico.
  + Diferncia entre CELLS y UNITS.
  + Comparar entre intervalos y analizar estadísticamente la volatilidad y definir la mejor. Comparar estrategias y en diferentes intervalos. Algo que sustente la selección de ese intervalo. Euristica debe tener. **likelihood**
* **Utilizar la auto-correlacion para análisis de intervalos**
* Utilizar Cross Validation (porcentajes de pruebas) en modelos.
* **Uso de PCT Change.**
  + Logaritmo de PCT Range e identificar diferencias.
  + Independiente del precio con PCT se puede predecir a donde se mueve.
* **Enfocarme en el análisis del dataset. Y Concluir sobre el análisis.**
* Avanzar mas con los modelos.
  + Implementar lo aprendido en el curso en mis Jupyter Notebooks.
* Probar no solo con el siguiente evento, sino varios adelante. Para LSTM u otros modelos.
* Hacer modelos y combinar Features.
* Se pretende tener una estrategia de trading que pueda tener reglas en base a un modelo o varios. (Ejemplo predecir por minutos, horas y días) en base a los resultados tomar una decisión).