**Especificación del modelo de ML**

1. Motivación del proyecto de ML

¿Cuál es el problema específico de ML que se quiere resolver?

El problema que se requiere resolver, es uno de aprendizaje supervisado, que pretende predecir en spread de crédito (serie de tiempo) que forman la curva de valoración de los TES UVR y la de bancos calificados AAA indexados al IPC.

¿Cómo este problema se articula con el problema de analítica formulado?

Teniendo en cuenta que se trata de una serie de tiempo, el problema será abordado con un modelo ARIMA, que involucra modelos Autorregresivos (AR) y de Media Móvil (MA), con un componente de integración que requiere que la serie sea estacionaria, para lo cual se deben calibrar el parámetro d. Lograr predecir la serie de spread crediticio bajo esta línea de trabajo, permitirá a la entidad conocer si debe emitir pronto y minimizar el costo de los recursos, o si en su lugar se debe dar espera y buscar un crédito de Tesorería.

1. Definición del problema

¿Cuál es la salida especifica que se va a predecir?

El spread de riesgo de crédito (diferencia entre la curva de valoración de un activo riesgoso y el activo libre de riesgo (TES: bonos emitidos por el gobierno)

¿Cuáles son los datos de entrada requeridos para el proceso de analítica?

Los datos de entrada corresponden a la misma serie, spread de crédito, histórica. Será un análisis univariado sobre el comportamiento de la serie.

Para el dataset a utilizar en ML describa:

1. Número de registros de entrenamiento

La base de datos contempla cinco años de historia. El entrenamiento se hará por medio de la validación cruzada, iniciando con dos años y agregando de a un día a la serie de predicción. La razón de iniciar con dos años radica en que este periodo incorpora un ciclo alcista y uno bajista de inflación que podrían ayudar a realizar una mejorar la predicción del modelo.

1. Existen datos no-balanceados, existen sesgos en los datos

No. La información a utilizar corresponde a información de mercado registrada en las plataformas de negociación.

1. ¿Qué tan frecuentemente se actualizan los datos?

Diariamente, dado que corresponde a datos del mercado.

¿Cuáles son los factores más relevantes que permitirían predecir la salida?

Teniendo en cuenta que se trabajará con un modelo univariado, los factores más relevantes estarán asociados a la correcta calibración de los modelos ARIMA.

1. Medidas de desempeño

¿Se tiene un marco de trabajo experimental de línea base?

Si:

¿Existe documentación?

Hay varios papers y bibliografía relacionada con la implementación de modelos de aprendizaje en serie financieras; sin embargo, el reto aquí está en la implementación en el mercado colombiano y en su aplicación para el problema relacionado con la fecha emisión de bonos o CDT´s.

¿Cuál es su desempeño y cómo se mide?

Dado que se trata de series de tiempo, las medidas de desempeño respecto a la precisión están asociadas a medias como el MSE o el RMSE. También se puede utilizar matrices de confusión para definir si la predicción respecto a la dirección de la serie es buena o no.

No:

¿Cuál es el marco experimental propuesto (que experimentos de predicción se requieren)?

¿Cómo se va a medir el desempeño del modelo de ML?

¿Cuál son las implicaciones de tener FP y FN (errores)?

¿Cuál es una buena solución?

¿Cuál es la literatura relacionada al problema (ej. papers y/o proyectos)?

1. Partición de los datos

¿Qué consideraciones se deben tener en cuenta para la partición de los datos (sesgos, imbalance de clases, múltiples etiquetas)?

Teniendo en cuenta que se trata de un problema de series de tiempo, la consideración a tener en cuenta respecto a la partición de los datos radica en considerar todo el periodo de tiempo a través de ventanas deslizantes.

¿Qué estrategia de partición se va a utilizar (random, LOOCV, K-folding)?

Cross Validation, Ventanas deslizantes.

1. Línea de tiempo

¿Cuándo es el deadline?

La serie tiene actualizaciones diarias, pero para efectos del ejercicio se tendrá en cuenta hasta el cierre de mes de junio de 2021. La fecha máxima de entrega será a inicio del mes de septiembre.

¿Cuándo se deben tener los primeros resultados?

Tan pronto se logre calibrar el modelo ARIMA, se espera empezar a tener resultados, es decir para la última semana de agosto de 2021.

¿Cuándo se debe tener la solución propuesta?

La solución deberá estar a inicios del mes de septiembre de 2021. Lo anterior teniendo en cuenta que la emisión de Bonos y CDT´s se espera hacer en el mes de Octubre.

1. Contactos

¿Quién es responsable del proyecto?

David del Valle, Director de Riesgo de Crédito, Mercado y Liquidez y Wilson Bohórquez, analista de riesgos de mercado y liquidez.

¿Quién es el responsable del acceso a los datos (en training y test)?

Analista de riesgos de mercado y liquidez.

¿Quién puede ayudar a entender el proceso, los resultados (experto de dominio)?

El equipo de investigaciones económicas podría ayudar con la interpretación de resultados o sugiriendo algunas variables macroeconómicas a incorporar.

1. Recursos para entrenamiento

¿Cuáles son los recursos computacionales (software/hardware) disponibles para entrenamiento?

El equipo de trabajo asignado por la compañía.

¿Cuál es el presupuesto (tiempo/dinero) disponible para entrenamiento?

No hay un presupuesto como tal, dado que los recursos que provee la empresa, son suficientes.

1. Despliegue

¿Cuáles son las restricciones tecnológicas para el despliegue del modelo (software particular, maquinas particulares)? En principio no hay restricciones.

¿Cuál va a ser el uso del modelo (consultas diarias, mensuales, etc)? A necesidad, cada vez que se requiera emitir CDT´s o bonos.

¿Cuál va a ser el mecanismo de acceso al modelo? Servidor Web a nombre de la Dirección de Riesgos.

¿Cómo se va a integrar el modelo con otros componentes del negocio (personas otros sistemas)?