Nombre: Credit Risk Assessment using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications

Abstract: Una estimación precisa del riesgo y su uso en modelos corporativos o globales de riesgo financiero, podría traducirse en un uso más eficiente de los recursos. Se hace un análisis comparativo de diferentes métodos de satistical y machine learning para clasificación en un conjunto de datos de préstamos hipotecarios con la motivación de comprender sus limitaciones y potencial.

Palabras Clave: Machine Learning, Deep Learning, hipotecas.

Presentación del Problema: Surge el interés de la evaluación del reisgo de crédito a partir de las crisis financieras que tuvieron lugar en las décadas de los 80´s y 90´s. Por lo que una mejor estimación de la precisión del riesgo podría ser traducida en un uso eficiente de los recursos. Entonces la idea crucial de esta investigación radica en encontrar predictores realistas del riesgo individual a través de modelos de riesgo realistas y precisos.

Presentación de la Metodología: La metodología consiste en hacer un análisis comparativo de diferentes métodos de clasificación en un conjunto de datos de préstamos hipotecarios de un gran banco comercial. Lo anterior se realiza mediante un estudio sistemático y una comparación con las técnicas tradicionales de clasificación estadística.

Se utiliza un enfoque de estrategia múltiple donde se aplican varios algoritmos con los mismos datos y se comparan sus resultados para encontrar el mejor modelo. Esto se justifica por el hecho de que es muy difícil seleccionar un modelo óptimo a priori sin conocer la complejidad real de un problema o conjunto de datos en particular. También se introduce una metodología específica para el análisis del modelo basado en el estudio de curvas de error para estimar el ruido / sesgo y la complejidad del modelo y el conjunto de datos.

El proceso o la construcción de un modelo y su aplicación a nuevos ejemplos de datos implican un costo computacional práctico. Esto debe

tenerse en cuenta, ya que puede limitar el tipo o los modelos que se pueden utilizar en una situación particular.

Se describen los elementos básicos de la metodología y el análisis de construcción de modelos que se emplean en los cuatro algoritmos considerados en el estudio. Los principales elementos de la metodología de análisis son:

· Exploración de parámetros del modelo básico.

· Análisis de importancia / sensibilidad de las variables.

· Análisis de errores de capacitación, prueba / generalización y evaluación, incluidas las matrices de rendimiento.

· Análisis de curvas de aprendizaje y estimaciones de ruido y parámetros de complejidad.

· Selección de modelo y combinación de resultados.

La Comisión Nacional de Bolsa de Valores ya había utilizado los datos para un modelo de regresión y, por lo tanto, requería poco procesamiento previo o manipulación antes de la construcción del modelo. Los datos consisten en un único conjunto de datos de 4.000 registros, cada uno de ellos correspondiente a una cuenta de cliente, y contiene un total de 24 atributos.

Los modelos logit y probit son similares, pero utilizan las distribuciones logísticas y normales acumulativas respectivamente. Una diferencia en estas distribuciones es que la distribución logística tiene colas más gruesas y esto a su vez produce pequeñas diferencias en el modelo, sin embargo, no hay bases teóricas para favorecer una técnica sobre la otra.

El modelo CART de Decision-Tree consiste en potentes modelos no paramétricos que producen predicciones precisas y reglas fácilmente interpretables para caracterizarlos. Son buenos representantes de la clase de algoritmos basados en reglas del árbol de decisiones.

Se eligio utilizar la arquitectura de red neuronal feedforward estándar compatible con el conjunto de herramientas de Darwin y se experimento

con varios algoritmos de entrenamiento: backpropagation, steepest descent, conjugate gradient, modified Newton, y genetic algorithm Los métodos de segundo orden, como el gradiente conjugado, permiten un entrenamiento mucho más rápido que la retropropagación estándar. También se investigó el efecto de cambiar las funciones de activación para la capa oculta: sigmoide, lineal e hipertangente. El algoritmo genético permite la optimización del peso en la región de la superficie de error, lo que podría ser difícil para los métodos basados en gradientes.

Resultados: Los resultados muestran que los modelos de árbol de decisión CART proporcionan la mejor estimación de incumplimiento con una tasa de error promedio de 8.31% para una muestra de entrenamiento de 2,000 registros. Como resultado del análisis de la curva de error para este modelo, concluimos que si hubiera más datos disponibles, aproximadamente 22,000 registros, se podría lograr una tasa de error potencial de 7.32%. Neural Networks proporcionó el segundo mejor resultado con un error promedio de 11.00%. El algoritmo K-Nearest Neighbour tenía una tasa de error promedio de 14.95%. Estos resultados superaron al algoritmo Probit estándar que alcanzó una tasa de error promedio del 15,13%.

Conclusiones: Si bien el articulo es algo longevo contribuye a la ciencia por que experimenta con varios modelos para un problema que sigue vigente hoy en dia, el problema de predecir si un clinte me pudede pagar el dinero que le preste. La combinación de diferentes estrategias y la aplicación de una metodología sistemática de construcción y selección de modelos ofrecen una perspectiva interesante para comprender las características y la utilidad de diferentes algoritmos o métodos de ajuste de datos

**

Nombre: Deep learning with long short-term memory networks for financial market predictions

Abstract: Las redes de larga memoria corta, son una herramienta novedosa para sequence leraning, son adecuadas en el ámbito de las series de tiempo financiera. Se implementan redes LSTM para predecir movimientos direccionales fuera de la muestra para las existencias constituyentes del S&P 500 desde 1992 hasta 2015. Específicamente, se encontró un patrón común entre las acciones seleccionadas para el comercio: exhiben una alta volatilidad y un perfil de rendimiento de reversión a corto plazo. Al final se utilizaron los resultados para diseñar la estrategia de supervisión a corto plazo para que dé cuenta de los retornos de LST.

Palabras Clave: Finanzas, statistical arbitrage, LSTM, machine learning, deep learning.

Presentación del Problema: el problema identificado consiste en que las tareas de predicción en series de tiempo financieras son notoriamente difíciles, principalmente impulsadas por el alto grado de ruido y la forma generalmente aceptada y semi-fuerte de eficiencia del mercado

Presentación de la Metodología: La metodología consiste en la aplicación de redes neuronales han establecido evidencia inicial de que son capaces de identificar estructuras (no lineales) en los datos del mercado financiero. En este contexto lo primero consiste en enfocarse en las redes de memoria a largo plazo (LSTM), una de las arquitecturas de aprendizaje profundo más avanzadas para tareas de aprendizaje secuencial

En segundo lugar, se persigue el objetivo de obtener resultados de la caja negra de las redes neuronales artificiales, revelando así las fuentes de rentabilidad. En tercer lugar, se sintetizan los hallazgos de la última parte en una estrategia comercial simplificada y basada en reglas que tienen como objetivo capturar lo más importante de los patrones sobre los que actúa LSTM para seleccionar acciones ganadoras y perdedoras.

Además, se aplicaron otros cinco pasos los cuales consisten en lo siguiente: Generación de training-trading set, definimos un período de estudio como un conjunto de capacitación y negociación, que consiste en un período de training de 750 días (aproximadamente tres años) y un período de trading de 250 días (aproximadamente un año). Dividimos todo nuestro conjunto de datos desde 1990 hasta 2015 en 23 de estos períodos de estudio con períodos de trading no superpuestos.

Generación de features y target, se generó el retorno de las secuencias para redes LSTM. Las redes LSTM requieren secuencias de características de entrada para el entrenamiento, es decir, los valores de las características en puntos consecutivos en el tiempo. La característica única es la devolución estandarizada de un día. Se optó por una secuencia de 240, que comprende la información de aproximadamente un año comercial.

En los targets, se definió un problema de clasificación binaria, es decir, la variable de respuesta Y st + 1 para cada stock s fecha t puede tomar dos valores diferentes.

Las redes LSTM pertenecen a la clase de redes neuronales recurrentes (RNN), es decir, redes neuronales cuya topología subyacente de conexiones interneuronales contiene al menos un ciclo. Las redes LSTM están compuestas por una capa de entrada, una o más capas ocultas y una capa de salida. El número de neuronas en la capa de entrada es igual al número de variables explicativas (espacio de características). El número de neuronas en la capa de salida refleja el espacio de salida, es decir, dos neuronas en nuestro caso que indican si un stock supera o no la mediana de la sección transversal en t + 1. La característica principal de las redes LSTM está contenida en la capa oculta (s) que consiste en las llamadas células de memoria.

Comparación de modelos se dio contrastando: random forest, red profunda y regresión logística, para comparar el LSTM, se eligió random forest, es decir, un método de aprendizaje automático robusto, pero de alto rendimiento, una red profunda estándar, es decir, para mostrar la

ventaja del LSTM, y una regresión logística, es decir, un clasificador estándar como línea de base.

Resultados: Los resultados se presentan en tres etapas. Primero, se analizan los retornos antes y después de los costos de transacción de 5 bps por media vuelta, y se contrasta el rendimiento de la red LSTM con el random forest, la red neuronal profunda y la regresión logística. En segundo lugar, se derivaron patrones comunes dentro de las acciones superiores y operativas, revelando así fuentes de rentabilidad. Tercero, se desarrolló una estrategia comercial simplificada basada en estos hallazgos, y se muestra que se puede replicar una parte del desempeño de LSTM con reglas transparentes, muchas de ellas basadas en anomalías tradicionales del mercado de capitales.

Conclusiones: Contribuye a la ciencia porque es un estudio reciente de la aplicación de redes neuronales recurrente a series de tiempo en el mercado financiero, lo cual viene a aportar a los análisis ya conocidos por los modelos econométricos de series de tiempo. En general, este documento realiza tres contribuciones clave a la literatura: la primera contribución se centra en la aplicación empírica a gran escala de las redes LSTM a las tareas de predicción de series de tiempo financieras.

En segundo lugar, se develo la caja negra \ LSTM, y dieron a conocer patrones comunes de acciones que se seleccionan para la negociación rentable. Se encontró que la cartera de LSTM consiste en acciones con un impulso por debajo del promedio, fuertes características de reversión a corto plazo, alta volatilidad y beta. Todos estos hallazgos se relacionan en cierta medida con anomalías existentes en el mercado de capitales.

Tercero, en base a los patrones comunes de la cartera de LSTM, se diseñó una estrategia comercial simplificada basada en reglas. Específicamente, redujimos ganadores de corto plazo compramos perdidas de corto plazo, y se mantuvo la posición por un día, al igual que en la aplicación LSTM. Con esta estrategia transparente y simplificada, se logran retornos de 0.23 por ciento por día antes de los costos de transacción. Cuando se regresa el rendimiento de la cartera LSTM a los de la estrategia simplificada, se observó una R al cuadrado de 52.1 por ciento para el

tramo largo y 53.6 por ciento para el tramo corto, lo que resulta en un alfa restante de la estrategia basada en LSTM de 31 bps por día antes de costos de transacción.

**

Nombre: Deep Learning for Mortgage Risk

Abstract: En este artículo se desarrolla un modelo de aprendizaje profundo sobre el riesgo hipotecario de varios períodos y luego se utiliza para analizar un conjunto de datos de origen y registros de rendimiento mensual sin precedentes para más de 120 millones de hipotecas originadas en los Estados Unidos entre 1995 y 2014. Tiene el alcance que los estimadores de estructuras de plazos de probabilidades condicionales de prepago, la ejecución hipotecaria y varios estados de morosidad incorporan la dinámica de un gran número de variables específicas del préstamo y macroeconómicas hasta el nivel del código postal.

Palabras Clave: Hipoteca, Machine Learning, Deep Learning, Mora.

Presentación del Problema: el problema es corresponde a que la literatura empírica sobre hipotecas identifica una serie de variables que ayudan a predecir el crédito hipotecario y el riesgo de pago anticipado, incluidos el puntaje e ingresos del crédito del prestatario, la relación préstamo-valor, la edad del préstamo, las tasas de interés y los precios de la vivienda. Para garantizar la capacidad de seguimiento econométrica, los investigadores a menudo imponen restricciones a los modelos empíricos que utilizan para estudiar el papel de varios factores de riesgo. Es importante destacar que la relación entre los factores y el rendimiento de la hipoteca generalmente se ve limitada a una forma predeterminada, con la opción estándar de ser lineal. Sin embargo, los datos de rendimiento de la hipoteca no respaldan tales restricciones.

Presentación de la Metodología: La metodología consiste en proponer un enfoque no lineal para abordar el problema ya descrito. Desarrollando un modelo de aprendizaje profundo de crédito hipotecario y riesgo de prepago en el que la relación entre los factores de riesgo y el rendimiento del préstamo no se basa en una forma pre especificada como en modelos empíricos anteriores. En nuestro enfoque, esta relación está completamente dictada por los propios datos, lo que minimiza la especificación errónea del modelo y el sesgo de las estimaciones de sensibilidad variable.

Los datos de la hipoteca fueron licenciados por CoreLogic, quien recopila los datos de los originadores y administradores de hipotecas. Es el conjunto de datos hipotecarios más completo estudiado hasta la fecha. Cubre aproximadamente el 70% de todas las hipotecas originadas en los EE. UU. Y contiene hipotecas de más de 30,000 códigos postales en los EE. UU. Las fechas de origen de las hipotecas van desde enero de 1995 hasta junio de 2014. El conjunto de datos incluye 25 millones de hipotecas de alto riesgo y 93 millones de hipotecas principales.9 Los datos del préstamo se dividen en (1) características del préstamo en el origen y (2) datos de rendimiento.

Cada hipoteca tiene una serie de características detalladas al inicio, como el puntaje FICO del prestatario, la relación préstamo-valor (LTV) original, la relación deuda-ingreso (DTI) original, saldo original, tasa de interés original, tipo de producto, tipo de propiedad, multas por pago anticipado, código postal, estado y muchos más. Muchas de las variables son categóricas con muchas categorías (algunas con hasta 20 categorías).

Se empleó el código postal de una hipoteca para hacer coincidir una hipoteca con factores locales como el precio mensual de la vivienda en ese código postal. Los precios de la vivienda se obtienen de Zillow y la Administración Federal de Vivienda (FHA). Los precios de las viviendas de Zillow están en el nivel de código postal de cinco dígitos y cubren aproximadamente 10,000 códigos postales. Para cubrir áreas menos pobladas que no están cubiertas por el conjunto de datos de Zillow, también incluimos precios de viviendas de la FHA que cubren todos los códigos postales de tres dígitos.

Descripción de la categorización de la deuda es la siguiente: Las hipotecas pueden realizar la transición entre 7 estados: actual, 30 días de atraso, 60 días de atraso, más de 90 días de atraso, ejecución hipotecaria, REO y pago. X días de morosidad simplemente significa que el prestatario de la hipoteca tiene X días de atraso en sus pagos al prestamista. Utilizamos el estándar establecido por la Asociación de Banqueros Hipotecarios de América para determinar el estado de morosidad. Se determina que una hipoteca está atrasada durante 1 mes si no se ha realizado el pago el último día del mes y el pago se debió el primer día del mes. REO significa

propiedad de bienes inmuebles. Cuando una hipoteca ejecutada no se vende en una subasta, el prestamista o administrador asumirá la propiedad de la propiedad. El pago puede ocurrir por una hipoteca prepaga, vencimiento, una venta corta o una hipoteca embargada que se vende en una subasta a un tercero (esto es nuevamente raro en comparación con los prepagos, que constituyen la mayor parte de los eventos pagados en el conjunto de datos).

Resultados: El resultado es que el modelo de aprendizaje profundo ajustado se utiliza para comprender la relación entre los factores explicativos y el comportamiento del prestatario. El análisis muestra que existen muchas relaciones altamente no lineales. Además, se encuentra que el comportamiento del prestatario tiene dependencias no triviales de la interacción no lineal entre múltiples factores.

Conclusiones: Contribuye a la ciencia inicialmente por que es de reciente lanzamiento y sirve como una herramienta adicional para el análisis del crédito hipotecario. Los hallazgos empíricos arrojan una gama de nuevas ideas importantes sobre el comportamiento de los prestatarios hipotecarios.

Se encontró que la relación entre el comportamiento del prestatario y los factores de riesgo es altamente no lineal, lo que cuestiona muchos modelos lineales estudiados en trabajos anteriores. Los efectos de interacción, donde el impacto de una variable depende de los valores de otras variables, son ubicuos.

Además, se encontró evidencia que sugiere que los prepagos son los más afectados; implican los efectos no lineales más fuertes entre todos los eventos. Los principales impulsores del pago anticipado, que incluyen saldos de préstamos pendientes originales y actuales, incluso antes de los factores estándar, como las tasas de interés y los diferenciales de tasas de interés, interactúan fuertemente. Los prestatarios Jumbo cuyo saldo pendiente actual es relativamente pequeño son los que tienen más probabilidades de pagar por adelantado.

Por último, los resultados empíricos tienen implicaciones significativas para los inversores de seguridad respaldada por hipotecas (MBS). El papel

dominante del desempleo resalta la exposición de los prestatarios a los ciclos económicos, que pueden ser una fuente sustancial de correlación de préstamo a préstamo. La prevalencia de esta correlación enfatiza la necesidad de que los inversores de MBS diversifiquen el riesgo hipotecario geográficamente, más allá de las características convencionales del prestatario resaltadas en la literatura.