Minería de datos y modelización predictiva

Profesor: Lorenzo Escot

Lorenzo Escot es Catedrático de Econometría en la Facultad de Estudios Estadísticos y codirector del Grupo de Investigación UCM "Análisis de Datos en Estudios Sociales, Género y Políticas de igualdad"

En las dos sesiones (de viernes y sábado) dentro del Módulo de Minería de Datos y Modelización Predictiva se presentarán **dos aplicaciones concretas de la modelización predictiva** en las que se utilizan muchas de las técnicas estadísticas de análisis de datos que ya se han visto previamente en el módulo. De hecho mi sesión es la última de modelización predictiva antes de comenzar con el módulo de *machine learning*, y de hecho se encuentra **entre los dos mundos**, el de la **modelización "inferencial" tradicional** (dónde el mejor modelo a emplear en nuestras predicciones es el que se ajusta mejor al tipo de variable dependiente y, sobre todo, a los supuestos que se hagan sobre la distribución del término de error) y la **modelización "algorítmica"** típica del ***machine learning*** (dónde los supuestos del término de error juegan un papel secundario y la competencia entre modelos es por saber cuál es el que predice mejor).

Más concretamente se utilizarán en las exposiciones de clase los modelos de regresión logística y los modelos de regresión lineal (con alguna variación). Sin embargo, todo lo que presente será también válido para ser aplicado con otros modelos predictivos basados en el aprendizaje automático que comenzaréis a ver cuando terminemos este módulo.

El contenido de esta semana está dividido en dos partes o sesiones, que se corresponden con las dos aplicaciones que vamos a ver:

- **Modelos de puntuación de Riesgo de Crédito (Scoring de riesgo de crédito)**
- **GIS y Modelos predictivos con Datos Espaciales.**

Todas las presentaciones, apuntes, materiales y scripts de Python estarán disponibles en la página web del máster y están organizados según estas dos aplicaciones: **Scoring** y **Espacia**l. A continuación paso a presentar los contenidos de cada una de estas dos aplicaciones

Sesión 1: Modelos de puntuación de Riesgo de Crédito (Scoring de riesgo de crédito)

1. **Introducción al concepto de Riesgo de Crédito** y a los modelos de Valoración o Puntuación del riesgo de crédito
2. Etapas en la construcción de una **tarjeta de puntuación de riesgo** (Scorecard)
    2.1. Selección y depuración de datos de clientes (datos faltantes, datos atípicos,...)
    2.2. Definición de la variable objetivo y la ventana temporal de observación de impagos
    2.3. Selección de variables explicativas mediante medidas de concentración: tramificación, agrupación, selección (WOE, Information Value y Gini) y transformación WOE de las variables predictoras
    2.4. Estimación y diagnosis de modelos de probabilidad: aplicación con regresión logística
3. El problema del sesgo de **selección muestral y la inferencia de denegados**
4. **Construcción de las Scorecard**, elección del umbral de riesgo óptimo y validación final del modelo

En primer lugar, veremos una aplicación de los modelos predictivos en la "**gestión de riesgos**", más concretamente el **Scoring o puntuación de Riesgo de Crédito**. Estas técnicas son aplicables también al cálculo de las Primas de Riesgo o; a los modelos para la **detección de Fraude**; los **modelos de fuga de clientes** (Customer Churn) ..... y en general a los modelos de **puntuación de riesgo de evento**.

Revisaremos las diferentes fases del análisis de riesgos, los métodos de estimación y diagnosis de los modelos de probabilidad; y repasaremos aspectos fundamentales a tener en cuenta en este tipo de análisis como son el **sesgo de selección muestral**, la **tramificación de variables continuas, la transformación WOE de variables** y los métodos de selección de variables explicativas utilizando cr**iterios de concentración**.

Detrás de estos modelos de puntuación del riesgo está la **regresión logística**, y como ya la habéis visto anteriormente yo hago sólo un recordatorio de estos modelos.

Así que me centro más en la **metodología de la construcción de los modelos de puntuación de riesgos**. Se trata de encontrar el mejor modelo para **separar a buenos y malos clientes**, a clientes que si les concedemos un préstamo nos lo devolverá sin problemas y a clientes que serán malos pagadores (incurrirán en **impago**). No queremos conceder préstamos a malos clientes, por eso se utiliza un modelo de puntuación del riesgo de crédito, para intentar predecir si un cliente al que tengo que conceder o no conceder un préstamo será o no buen pagador.

Nos centraremos en la **puntuación de riesgo de nuevos clientes** (**scoring de admisión**). Individuos que nos solicitan un préstamo pero de los que no tenemos ninguna información, sólo la que le preguntemos a ellos en el momento para decidir si le concedemos o no el crédito.

Como digo, **detrás de la puntuación del riesgo hay un modelo de regresión**, un modelo de probabilidad, la logística, o cualquier modelo que sirva para estimar probabilidades de que suceda algo: en nuestro caso el impago (también valdrían árboles, redes, etc… o cualquiera de los denominados **modelos de clasificación** binaria) . Dicho modelo de probabilidad utilizará una serie de variables sobre el nuevo cliente que ayuden a predecir su probabilidad de impago.

Una cuestión importante es que normalmente de los nuevos clientes no tengo información, por eso tengo que preguntarle directamente al nuevo potencial cliente por las variables que necesito para evaluar su riesgo. Por eso es fundamental elegir **pocas variables predictivas que ayuden mucho a separar a los buenos de los malos clientes**.

Se presentarán diferentes fases del análisis de riesgos; los métodos de diagnosis de los modelos de probabilidad; y se prestará especial atención a aspectos fundamentales a tener en cuenta en este tipo de análisis, como son el sesgo de selección muestral y la inferencia de rechazados, la tramificación de variables continuas, la transformación WOE de variables categóricas y los métodos de selección de variables explicativas utilizando criterios de concentración. Presentaré en este sentido los **WoE y el Information Value**, que son una medida basada en la concentración de los malos clientes en determinadas categorías de las variables explicativas. Estos criterios de concentración ayudan a hacer una selección inicial de variables candidatas a formar parte del modelo de predicción de la probabilidad de impago, esto es de la batería de preguntas que debemos hacer a los potenciales nuevos clientes que vienen a solicitarnos un crédito.

Desarrollo de las clases: durante la clase se realizará primero una presentación teórica de los contenidos y posteriormente una aplicación práctica de scoring de clientes de tarjetas de crédito utilizando, fundamentalmente, el paquete **OptBinning** de python.

Para seguir la práctica utilizaremos Anaconda para crear primero un entorno de desarrollo para esta práctica donde instalaremos el paquete OptBinning (también puede seguirse toda la práctica utilizando Google Colab). Después la práctica la seguimos con un archivo jupyter notebook (.ipynb) que se pondrá previamente a disposición de los estudiantes (aunque también presentaremos las ayudas que ofrece VS Code para desarrollar este código).

Aunque la librería central que utilizo es OptBinning, **las librerías que utilizo en la práctica son**:

- numpy
- pandas
- matplotlib.pyplot (y también seaborn)
- scipy.stats
- Scikit-learn
- optbinning

Documentación Adjunta

documentación en el campus virtual:

- *Apuntes de Scoring de Riesgo De Crédito.pdf*: Son unos apuntes que he preparado con los contenidos de la sesión (como los he preparado yo os diré que están muy bien, por lo que se agradecen comentarios para su mejora)
- *CreditScoring_BigDataUCM.pdf:* Es la presentación de diapositivas que utilizaré en mis clases.
- Naeem Siddiqi.pdf: Es un manual de referencia básico, no muy técnico, para que el que lo necesite profundice en todo el proceso de construcción de tarjetas de puntuación
- datos *germancredit.csv* que se utiliza en la práctica de clase y que tendréis que cargar para poder realizarla
- Script *Clase_scorecard.ipynb* un jupiter notebook que contiene la práctica que desarrollaremos en clase

Prueba de Evaluación: consistirá en una serie de preguntas tipo test sobre la construcción de modelos de Scoring, con preguntas de tipo teórico y preguntas aplicadas utilizando el paquete optbinning según lo visto en la clase. Se proporciona un conjunto de datos para que los estudiantes realicen un modelo de riesgo y hagan una valoración de un conjunto de 32 nuevos clientes solicitantes de una tarjeta de crédito (pregunta 9 del cuestionario). Toda la tarea se entregará a través de la plataforma Moodle (el campus virtual del máster).

Tiempo aproximado de realización de la tarea 2 horas.

Bibliografía básica:

- Anderson, R(2007) *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation* . Oxford University Press
- Bolder, J. D. (2019). Credit-Risk Modelling: Theoretical Foundations, Diagnostic Tools, Practical Examples, and Numerical Recipes in Python. Springer.
- Escot, L (2024): Apuntes de modelización del Riesgo Crediticio
- Mays,E and Niall Lynas (2011) Credit Scoring for Risk Managers: The Handbook for Lenders.Createspace (ISBN13: 9781450578967)
- Siddiqi, N. (2006): Credit Risk Scorecards. Devoloping and implementing Intelligent Credit Scoring. J Wiley & Sons
- Trueck, S, & Rachev, Svetlozar (2009): Rating Based Modeling of Credit Risk. Theory and Application of Migration Matrices. Elsevier

Sesión 2. GIS y Modelos predictivos con datos espaciales

1. **Introducción a los Sistemas de Información Geográfica (GIS) y representación de datos georeferenciados**
    - 1.1. Mapas, escalas, proyecciones y sistemas de coordenadas
    - 1.2. Cartografías (shapes) con Geopandas
    - 1.3. Tipos de datos Espaciales o georeferenciados: datos Raster y datos Vectoriales (puntos, líneas y polígonos)
    - 1.4. Ejemplo: los mapas de coropletas con geopandas (Choropleth maps)
    - 1.5. Ejemplo: uso de OpenStreetMaps y de Leaflet a través de geopandas
    - 1.6. Ejemplo: uso de las Apis de OpenStreetMaps para el cálculo de rutas óptimas
2. **Midiendo la Dependencia Espacial**
    - 2.1. La Ley de Tobler o "Primera Ley de la Geografía"
    - 2.2. Heterogeneidad Espacial vs Dependencia Espacial
    - 2.3. Distribuciones puramente aleatorias vs patrones de dependencia espacial
    - 2.4. Matriz de Vecindad o pesos espaciales: contigüidad y distancias
    - 2.5. El operador Retardo Espacial
3. **Análisis Exploratorio de Datos Espaciales**
    - 3.1. Midiendo la dependencia espacial: estadísticos locales y globales
    - 3.2. ***I*** de Moran, ***C*** de Geary, ***G(d)*** de Getis y Ord.
    - 3.3. El gráfico de Moran y contrastes de significatividad de ausencia de correlación espacial
    - 3.4. Agrupación espacial y mapas LISA (Local Indicator of Spatial Association)
4. **Modelos de Econometría Espacial: Especificación y diagnosis**
    - 4.1. Modelo de Retardo Espacial
    - 4.2. Modelo de Error Espacial
    - 4.3. Modelo de Durbin
    - 4.4. Otras especificaciones
    - 4.5. Diagnosis de los Errores
    - 4.6. Efectos directos y efectos indirectos

En la segunda de las aplicaciones dejamos los "riesgos" y nos adentraremos en el apasionante mundo de la "estadística espacial". Veremos qué tienen de particular los **datos espaciales**, qué son los **Sistemas de Información Geográfica** (GIS), y presentaremos algunas herramientas para la elaboración de mapas y el análisis de la distribución geográfica de datos espaciales. A partir de ahí nos centraremos en el análisis de la **autocorrelación espacial** y en los **modelos de econometría espacial.**

Los **datos espaciales** son datos que incluyen un índice o referencia de su localización. Hablaremos primero de forma muy resumida de los diferentes sistemas de información Geográfica GIS, sistemas que son necesarios para hacer una representación de los datos espaciales geo-referenciados...Y ya más adelante introduciremos el concepto de la autocorrelación o dependencia espacial, fundamental en todo el análisis de datos espaciales.

A modo de resumen, los modelos de regresión espacial son modelos de regresión con una variable dependiente y un conjunto de variables independientes o explicativas, entre las que se incluye una variable explicativa con la que se trata de captar la dependencia espacial, es decir, la dependencia con la situación en un entorno cercano o próximo a cada objeto espacial.... ya los vamos comentando.

Desarrollo de las clases: durante la clase se realizará primero una presentación teórica de los contenidos y después una aplicación práctica utilizando Python, fundamentalmente las librerías **Geopandas** y **Pysal.**

Para seguir la práctica se utilizará Anconda para crear primero un entorno de desarrollo para esta práctica donde instalaremos los paquetes Geopandas y Pysal (también puede seguirse las prácticas utilizando Google Colab). Después la práctica la seguimos con archivos jupyter notebook (.ipynb) que se pondrán previamente a disposición de los estudiantes. Más concretamente las librerías utilizadas son:

- Pandas
- matplotlib.pyplot
- seaborn
- geopandas
- Pysal

Documentación Adjunta

- *Econometría Espacial_BigData_Python.pdf*: es el archivo pdf con la presentación que utilizo en la clase
- *Bibliografía Espacial7z.7z* : archivo (comprimido con 7zip) que contiene un artículo de Ignacio Alonso Fernández-Coppel que a mí me ayudó en su día a entender que eran los GIS y los diferentes sistemas de proyección. También hay un artículo breve que describe los principales modelos de econometría espacial y un excelente manual para el análisis de datos espaciales
- fichero comprimido *cartografíasPython.7z* Es importante que los descomprimáis respetando la estructura de carpetas, una de cartografías y otra de datos, porque luego hago referencia a ellas dentro de las prácticas.
- Script (jupiter notebook) *1_1_IntroRepresentación Espacial*: donde os muestro como cargar cartografías y realizar mapas de coropletas fácilmente utilizando **geopandas**
- Script (jupiter notebook) *1_2_Geocodificacion.ipynb*: donde os muestro como geolocalizar objetos a partir de su dirección postal utilizando la API de Openstreetmap utilizando **geopandas**
- Script (jupiter notebook) *2_Regresión Espacial.ipynb*: aquí os muestro como cuantificar la auto correlación espacial y como incluirla en vuestros modelos predictivos (yo me limito a verlo en un modelo de regresión lineal sencillo) utilizando **Pysal**

Prueba de Evaluación: consistirá en la realización de una serie de tareas tipo test en el que los alumnos tendrán que aplicar el contenido de la clase (creación de mapas, estadística descriptiva espacial, y la especificación, estimación y diagnosis de un

Modelos Espaciales) utilizando Python. Dichos tests se relizarán a través de la plataforma Moodle.
Tiempo aproximado de realización de la tarea 2 horas.

Bibliografía básica:

- Anselin y Rey (2014): *Modern Spatial Econometrics in Practice,* GEODA Press
- Anselin, L. (1988). *Spatial Econometrics: Methods and Models. Kluwer Academic Press.*
- Bonny P. McClain (2022). Python for Geospatial Data Analysis: Theory, Tools and Practice for Location Intelligence. Ed O'Reilly
- Dubé, J. y Legros, D. (2014): *Spatial Econometrics Using Microdata*. Editorial Iste Ltd and John Wiley & Sons.
- *Elhorst, J.P. (2010). Applied Spatial Econometrics: Raising the Bar, Spatial Economic Analysis, 5(1), 9-28.*
- *Elhorst, J.P. (2014). Spatial Econometrics: From Cross-Sectional Data to spatial Panels. Springer*
- LeSage, J., & Pace, R. K. (2010). *Introduction to Spatial Econometrics*. CRC Press.
- Sergio J. Rey; Dani Arribas-Bel; y Levi J. Wolf (2020): Geographic Data Science with Python. https://geographicdata.science/book/intro.html