Primera Entrega de Proyecto

Profesor:

RAÚL RAMOS POLLÁN

Materia:

Introducción a la Inteligencia Artificial

Estudiantes:

César Augusto López Castillo

Mateo Yepes Sierra

William Alexander Torres Zambrano

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN

JULIO 2002

1) **Planteamiento del problema**

Si se le pide a un comprador que describa la casa de sus sueños, probablemente no empezará por la altura del techo del sótano o la proximidad a una vía férrea este-oeste. Pero el conjunto de datos de este concurso demuestra que hay muchas más cosas que influyen en las negociaciones sobre el precio que el número de dormitorios o una valla de malla blanca.

Con 79 variables explicativas que describen (casi) todos los aspectos de las viviendas residenciales en Ames, Iowa, esta competición le reta a predecir el precio final de cada vivienda.

El **problema consiste** en predecir el precio de venta de cada casa. Para cada ID del conjunto de pruebas, se debe predecir el valor de la variable **SalePrice**.

2) **Dataset**

Vamos a usar el dataset de Kaggle perteneciente a la competencia "House Prices - Advanced Regression Techniques":

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

Descripción de los archivos:

- train.csv - Set de entrenamiento
- test.csv - Set de testeo
- data_description.txt - Descripción completa de cada columna, preparada originalmente por Dean De Cock pero ligeramente editada para que coincida con los nombres de las columnas utilizados aquí
- sample_submission.csv - Una presentación de referencia a partir de una regresión lineal sobre el año y el mes de la venta, los metros cuadrados del lote y el número de habitaciones

Descripción de los datos:

El los archivos de datos encontraremos las siguientes variables:

- SalePrice - El precio de la propiedad en dólares. Esta es la variable objetivo que vamos a predecir
- MSSubClass: La clase de edificación
- MSZoning: la clasificación de la zona general
- LotFrontage: distancia lineal en pies que está conectada a la carretera
- LotArea: área del lote en pies al cuadrado
- Street: Tipo de carretera de acceso
- Alley: tipo de callejón para acceder
- LotShape: forma general de la propiedad
- LandContour: llanura de la propiedad
- Utilities: tipo de utilidades disponibles

- LotConfig: configuración del lote
- LandSlope: pendiente de la propiedad
- Neighborhood: ubicaciones físicas dentro de los límites de la ciudad de Ames
- Condition1: Proximidad de la carretera principal o vía férrea
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: tipo de vivienda
- HouseStyle: estilo de la vivienda
- OverallQual:Material general y calidad del acabado
- OverallCond: calificación de la condición general
- YearBuilt: fecha de la construcción original
- YearRemodAdd: fecha de remodelación
- RoofStyle: Tipo de tejado
- RoofMatl:material del tejado
- Exterior1st: Revestimiento general de la casa
- Exterior2nd: Revestimiento general de la casa si hay más de un material
- MasVnrType: tipo de chapa de mampostería
- MasVnrArea: area de revestimiento de mampostería en pies cuadrados
- ExterQual: calidad del material exterior
- ExterCond: condición actual del material en el exterior
- Foundation: tipo de cimiento
- BsmtQual: altura del sótano
- BsmtCond:condición general del sótano
- BsmtExposure: cantidad de muros en el sótano
- BsmtFinType1: calidad del área terminada del sótano
- BsmtFinSF1: Pies cuadrados terminados de tipo 1
- BsmtFinType2: Calidad de la segunda área terminada (si existe)
- BsmtFinSF2: Pies cuadrados terminados de tipo 2
- BsmtUnfSF: Pies cuadrados de superficie de sótano sin terminar
- TotalBsmtSF: Total de pies cuadrados de superficie del sótano
- Heating: Tipo de calentamiento
- HeatingQC: Calidad y estado de la calefacción
- CentralAir: Aire acondicionado central
- Electrical: Sistema eléctrico
- 1stFlrSF: Pies cuadrados del primer piso
- 2ndFlrSF: Pies cuadrados del segundo piso
- LowQualFinSF:Pies cuadrados de calidad inferior (todas las plantas)
- GrLivArea: Superficie habitable por encima del nivel del suelo (pies cuadrados)
- BsmtFullBath: Baños completos en el sótano
- BsmtHalfBath: Medios baños en el sótano
- FullBath: Baños completos sobre el nivel del suelo
- HalfBath: Medios baños sobre el nivel del suelo
- Bedroom: Número de habitaciones por encima del nivel del sótano
- Kitchen: Número de cocinas
- KitchenQual: Calidad de la cocina
- TotRmsAbvGrd: Total de habitaciones sobre el nivel del suelo (no incluye los baños)

- Functional: Valoración de la funcionalidad del hogar
- Fireplaces: número de chineneas
- FireplaceQu: Calidad de la chimenea
- GarageType: Localización del garaje
- GarageYrBlt: Año de construcción del garaje
- GarageFinish: Acabado interior del garaje
- GarageCars: Tamaño del garaje en capacidad de carros
- GarageArea: Tamaño del garaje en pies cuadrados
- GarageQual: Calidad del garaje
- GarageCond: Condición del garaje
- PavedDrive: Calzada pavimentada
- WoodDeckSF: Superficie de la cubierta de madera en pies cuadrados
- OpenPorchSF: Área del pórtico abierto en pies cuadrados
- EnclosedPorch: Superficie del pórtico cerrado en pies cuadrados
- 3SsnPorch: Superficie del pórtico de tres estaciones en pies cuadrados
- ScreenPorch: Superficie del pórtico en pies cuadrados
- PoolArea: Área de la piscina en pies cuadrados
- PoolQC: Calidad de la piscina
- Fence: Calidad de valla
- MiscFeature: Características diversas no incluidas en otras categorías
- MiscVal: $Valor de la función miscelánea
- MoSold: Mes de vendido
- YrSold: Año de vendido
- SaleType: Tipo de venta
- SaleCondition: Condición de venta

3) Métrica de Desempeño

La métrica para medir el desempeño de la predicción será con el error cuadrático medio (RMSE) entre el logaritmo del valor predicho y el logaritmo del precio de venta observado.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Donde:

N: es el número de datos totales

Predicted: los datos que predice el modelo

Actual: el valor real del dato

En cuanto a la métrica de negocio, se busca predecir el valor de la casa a partir de ciertas características que la describen. Se espera que esto ayude a ahorrar tiempo y dinero en procesos de selección por parte de los clientes.

4) Desempeño

Lo que se espera de este modelo es la predicción de costo de cada una de las casas de Ames, Iowa, basados en varias características que se proporcionarán a gusto del cliente. Se espera entonces que éstas predicciones mejoren los análisis financieros de clientes y organizaciones de bienes raíces. Finalmente se puede aprovechar esta información para mejorar modelos de predicción de este tipo, determinar cuánto se está ahorrando en costos y también evaluar el costo de casas para la venta.

5) Referencias bibliográficas

House Prices - Advanced Regression Techniques | *Kaggle*. (2016). Kaggle. https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview/description