# DS Onboarding Evaluation

Objective: To create a ML model to predict house value for properties in Los Angeles county.

Problem Statement

Our client is a global real-estate company, XBRE. They provide property valuation services to homeowners, who want to sell or mortgage their houses. They are having troubles handling the volume of valuation requests in the given time frame. They are also facing legal issues as they are unable to explain the value drivers for the valuation in an objective manner.

They have therefore reached out to Merilytics to build an Automated Valuation model to predict the valuation for properties. Since the client has high footprint in CA, Merilytics will do a pilot project for AVM model in Los Angeles area.

Based on our discussions with the VP-Data Analytics, the project blueprint has been approved. Merilytics would be responsible for the entire project – starting from identifying the relevant data sources for enrichment to validating the model.

The pilot project and report will follow the CRISP-DM framework for exploration and modeling. We are sharing a pic of the steps in the process and activities involved in each step.  
Table

Description automatically generated

Based on internal discussions, we have come up with the following steps to tackle the project,

* Use the raw data containing property features for predicting *Total Value* of the property. ***We have shared a data dictionary for the property data, along with the list of columns which can not be used for modeling purposes as those features would not be available for test data***. (Refer Appendix)
* Perform exploratory analysis on various features to understand the distribution and interactions. Use appropriate graphs for exploration.
* Use US Census API to enrich the dataset with appropriate demographic variables. Some variables that could be used for the predictive model are ***Total Population, Total Housing units, Median age of population, Average household size, Occupancy status, etc***
* Transform and process the features appropriately, providing adequate reasoning behind the same.
* Prepare and evaluate predictive models to adequately predict valuation of properties (***Total Value column in data***). Explore a mix of linear and non-linear machine learning and deep learning models for the modeling exercise.
  + Identify appropriate metrics to evaluate the models. Use cross-validation or appropriate validation techniques to evaluate the model. Perform hyper-parameter tuning to identify the best model to predict valuation of properties.
* Prepare a report based on your findings. Report should contain all the assumptions you used while preparing the model, the summary of **key insights and value drivers (key features such as feature importance, SHAP, etc)** identified for the property. Present the insights with appropriate visualizations.

Census API details

* Census API should be used for data enrichment by leveraging demographic information.
* You could use the **Census Geocoding API** for getting county, census tract/block information for each address.
  + **Documentation:** <https://www.census.gov/data/developers/data-sets/Geocoding-services.html>
  + **Examples:** <https://geocoding.geo.census.gov/geocoder/Geocoding_Services_API.pdf>
* **Documentation**: <https://www.census.gov/data/developers/data-sets/TIGERweb-map-service.html>; <https://api.census.gov/data/2010/dec.html>
* API endpoints: <https://api.census.gov/data/2010/dec/sf1>
  + To get a key, simply register with your mail id. You will get a key in mail within 10 mins.
* **Use the Decennial SF1 dataset.**

Guidelines

* Code should be production-level, modular and well-commented. Adhere to PEP8 guidelines as far as possible.
* Ensure code is efficient and uses pipelines, wherever possible.

Evaluation points

* **Modeling artifacts** –
  + Code files related to exploration, modeling and evaluation.
  + Model pickles
* **Documentation --** Documentation should adhere to CRISP-DM framework. It should be self-sufficient for client and Merilytics team members to understand the modeling exercise.
* **Power Point** – Summarize the model as a presentation for the C-Suite executives. **(Only for Sr. Associate, Manager)**

Instructors/Evaluators

* To be assigned by the Onboarding team. Setup milestone meetings with your instructor after exploring every stage of CRISP-DM framework.

**Appendix**

**Data Dictionary**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Type** |
| ZIPcode | Zip code in the situs\* address, either 9-digit or 5-digit. \*Situs pertains to the parcel’s location and almost always matches the actual street address. | Plain Text |
| TaxRateArea\_CITY | Incorporated city name based on tax rate area. | Plain Text |
| AIN | Unique 10-digit number assigned to each parcel within Los Angeles County. This is the same as AssessorID but here stored without the imbedded hyphens (and as a number rather than text). Also known as APN. | Number |
| RollYear | Assessment roll year. Properties are assessed as of the January 1 valuation date (lien date) for enrollment in July and property tax billing in October. | Number |
| TaxRateArea | Code that specifies the geographic area used in establishing the tax rates applied to the property. | Plain Text |
| AssessorID | Unique 10-digit number assigned to each individual parcel in Los Angeles County. This Assessor Identification No. is also known as Assessor Parcel No | Plain Text |
| PropertyLocation | Situs\* address of the property (HouseNo HouseFraction StreetDirection StreetName UnitNo City ZIPcode5). \*Situs pertains to the parcel’s location and almost always matches the actual street address. For incorporated cities, City here should be that city name; for unincorporated areas, City may be the name of the nearest incorporated city or a community name. | Plain Text |
| PropertyType | Abbreviated form of property use type based on PropertyUseCode: SFR=Single Family Residence CND=Condominium R-I=Residential-Income C/I=Commercial/Industrial VAC=Vacant OTH=Other | Plain Text |
| PropertyUseCode | 4-character property use code describing the overall use of the property. Breakdown by character is included in this dataset. | Plain Text |
| GeneralUseType | General use type of the property (1st character of PropertyUseCode). | Plain Text |
| SpecificUseType | More specific use type of the property (2nd character of PropertyUseCode). | Plain Text |
| SpecificUseDetail1 | Additional property usage detail (3rd character of PropertyUseCode). | Plain Text |
| SpecificUseDetail2 | Additional property usage detail (4th character of PropertyUseCode). | Plain Text |
| totBuildingDataLines | Total number of building data lines (individual structures). Although the Assessor maintains a complete set of property records on paper within the parcel jackets, only up to 5 building data lines can be stored in the current property data base. The 5 most significant structures are typically the ones selected. | Number |
| YearBuilt | Year property was originally built, If more than one structure (totBuldingDataLines>1) this is will be the earliest year built. | Number |
| EffectiveYearBuilt | Effective year built taking into account subsequent construction, remodeling, building maintenance, etc. If more than one structure (totBuldingDataLines>1) this is the latest effective year built. | Number |
| SQFTmain | Total square footage of the main structure(s), excluding garage area, porch, enclosed patio, etc. | Number |
| Bedrooms | Total number of bedrooms. | Number |
| Bathrooms | Total number of bathrooms, whether quarter, half, three-quarter or full bath. | Number |
| Units | Total number of living units. | Number |
| RecordingDate | Latest recording date, whether reappraisable or non-reappraisable. May include fictitious dates used in special processing (such as 19670245 or 19870633) and incomplete dates (such as 19790000). | Plain Text |
| LandValue | Land value on this assessment roll. | Number |
| LandBaseYear | Land base year established by Proposition 13. Changes to land base year are triggered only by re-appraisable change-of-ownership. | Number |
| ImprovementValue | Improvement value on this assessment roll. | Number |
| ImpBaseYear | Improvement base year established by Proposition 13. Changes to improvement base year are triggered only by re-appraisable change-of-ownership or major new construction. | Number |
| TotalLandImpValue | Total LandValue + ImprovementValue on this assessment roll. | Number |
| HomeownersExemption | Homeowner's exemption value on the assessment roll. | Number |
| RealEstateExemption | Real estate exemption value on this assessment roll (Church, Religious, Welfare or Veteran; full or partial) | Number |
| FixtureValue | Fixture value on this assessment roll. | Number |
| FixtureExemption | Fixture exemption value on this assessment roll. | Number |
| PersonalPropertyValue | Personal property value on this assessment roll. | Number |
| PersonalPropertyExemption | Personal property exemption value on this assessment roll. | Number |
| isTaxableParcel? | Y=taxable fee parcel; N=non-taxable non-fee parcel. Non-taxable non-fee parcels are identified by having 8th digit of AssessorID either 3, 8, or 9 (i.e. last 3 digits of AssessorID either 3xx, 8xx or 9xx). Assessed values for non-taxable parcels are relatively meaningless and so to avoid inadvertent comingling with taxable values, the assessed values and exemption values for NON-TAXABLE parcels have here been ZEROED-OUT. Also, because of lesser priority given to maintaining their property records, the parcel usage, building information, and even the property address shown here for non-taxable parcels may not be up-to-date. | Plain Text |
| TotalValue | Total value = LandValue + ImprovementValue + FixtureValue + PersonalPropertyValue Note: Assessed values for non-taxable parcels (isTaxableParcel?='N') are relatively meaningless and so to avoid inadvertent comingling with taxable values, the assessed values and exemption values for NON-TAXABLE parcels have here been ZEROED-OUT. | Number |
| TotalExemption | Total exemption value = HomeownersExemption + RealEstateExemption + FixtureExemption + PersonalPropertyExemption Note: Assessed values for non-taxable parcels (isTaxableParcel?='N') are relatively meaningless and so to avoid inadvertent comingling with taxable values, the assessed values and exemption values for NON-TAXABLE parcels have here been ZEROED-OUT. | Number |
| netTaxableValue | Net taxable value = Total value minus Exemption value (LandValue + ImprovementValue + FixtureValue + PersonalPropertyValue - HomeownersExemption - RealEstateExemption - FixtureExemption - PersonalPropertyExemption). Note: Assessed values for non-taxable parcels (isTaxableParcel?='N') are relatively meaningless and so to avoid inadvertent comingling with taxable values, the assessed values and exemption values for NON-TAXABLE parcels have here been ZEROED-OUT. | Number |
| SpecialParcelClassification | Special parcel classification, such as type of non-taxable parcel. Informational only. | Plain Text |
| AdministrativeRegion | Assessor's administrative office responsible for maintaining the property records for this parcel. Parcels are assigned based on type of property and geographic location. | Plain Text |
| Cluster | Cluster codes describe a geographical area by which similar types and uses of parcels are grouped. For residential and smaller commercial/industrial properties, clusters generally identify market areas. For more complex properties,clusters may only be used for internal work flow and control. | Plain Text |
| ParcelBoundaryDescription | Although often an abbreviated version of the recorded legal description, the parcel boundary description is NOT a legal description. It is an internal description used by the Assessor for assessment purposes. | Plain Text |
| HouseNo | House no. in the situs\* address. If a property has more than one street address, the situs address is typically (but not always) the one with the lowest street no. \*Situs pertains to the parcel’s location and almost always matches the actual street address. | Number |
| HouseFraction | Fraction associated with HouseNo in the situs\* address (example 1/2). \*Situs pertains to the parcel’s location and almost always matches the actual street address. | Plain Text |
| StreetDirection | Direction associated with the StreetName in the situs\* address : N, S, E, W, or blank. \*Situs pertains to the parcel’s location and almost always matches the actual street address. | Plain Text |
| StreetName | Street name in the situs\* address. \*Situs pertains to the parcel’s location and almost always matches the actual street address. | Plain Text |
| UnitNo | Unit no. in the situs\* address. \*Situs pertains to the parcel’s location and almost always matches the actual street address. | Plain Text |
| City | City name and State in the situs\* address. For incorporated cities, this should be that city name; for unincorporated areas this may be the name of the nearest incorporated city or a community name. \*Situs pertains to the parcel’s location and almost always matches the actual street address. | Plain Text |
| ZIPcode5 | 5-digit zip code of the situs\* address. \*Situs pertains to the parcel’s location and almost always matches the actual street address. | Number |
| rowID | Simple row identifier comprised of concatenated RollYear and AIN. | Plain Text |
| CENTER\_LAT | Geographic latitude of the parcel. | Number |
| CENTER\_LON | Geographic longitude of the parcel. | Number |
| Location 1 |  | Location |

**Columns Not to be Used in Modeling**

|  |
| --- |
| **Columns Not to be used for Modeling** |
| LandValue |
| LandBaseYear |
| ImprovementValue |
| ImpBaseYear |
| TotalLandImpValue |
| HomeownersExemption |
| RealEstateExemption |
| FixtureValue |
| FixtureExemption |
| PersonalPropertyValue |
| PersonalPropertyExemption |
| TotalExemption |
| netTaxableValue |