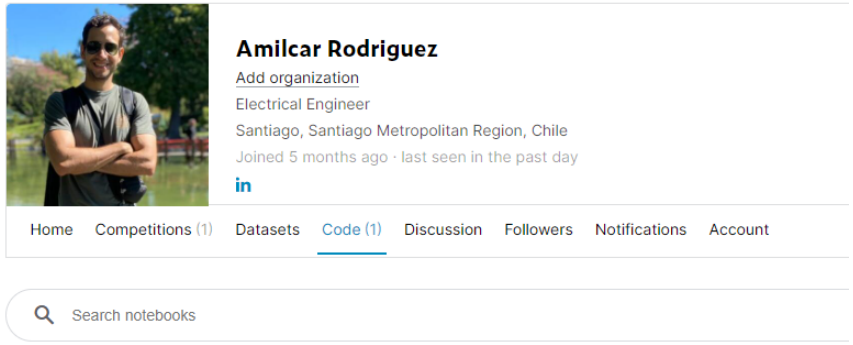
# Predicción de costo de venta de una propiedad

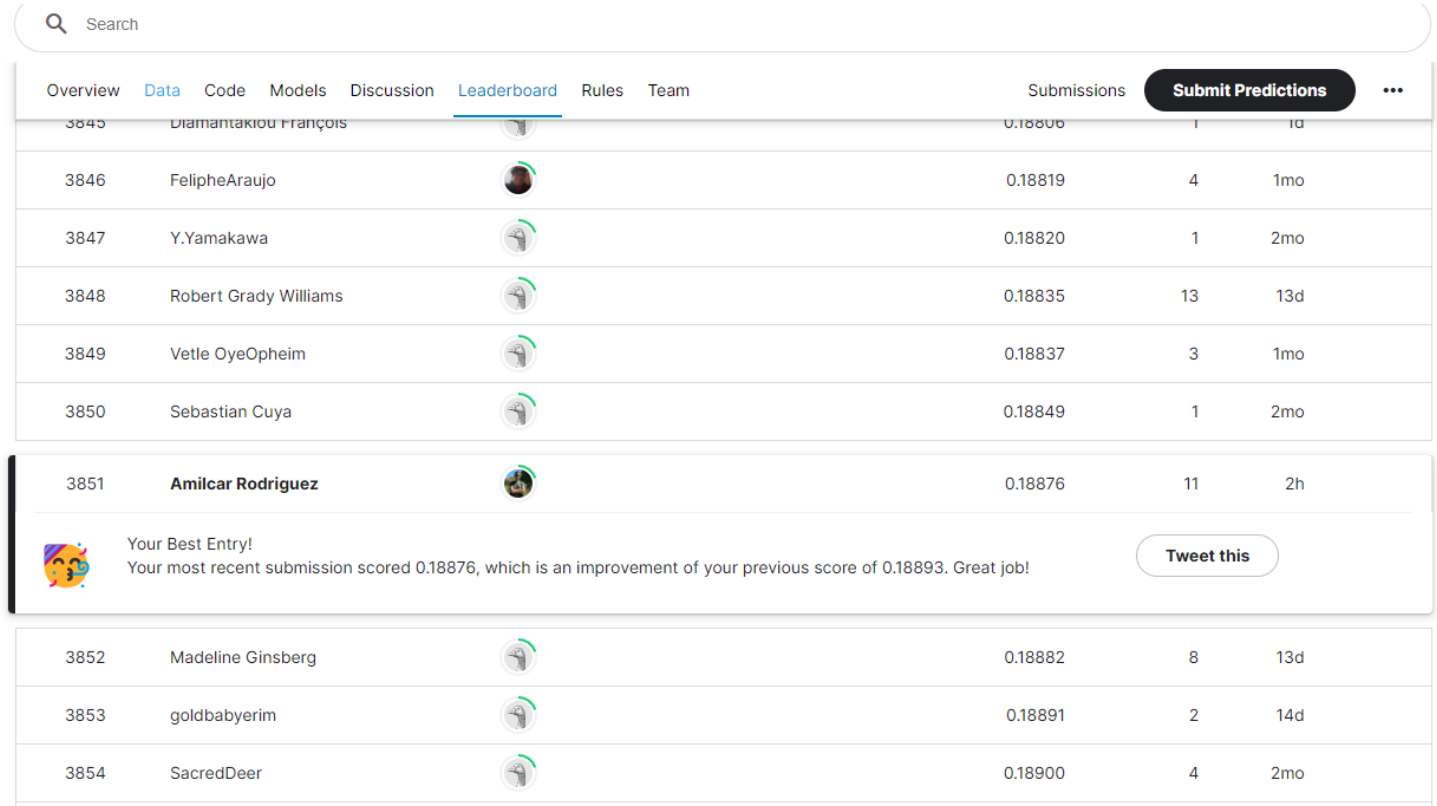
# RESUMEN DEL TRABAJO

## Perfil Kaggle

https://www.kaggle.com/amilcarrodriguez




## Resultados Kaggle




## Integrantes

| | |
|---|---|
| Amilcar Rodriguez | joserba91@gmail.com |

# DESCRIPCIÓN DEL PROBLEMA

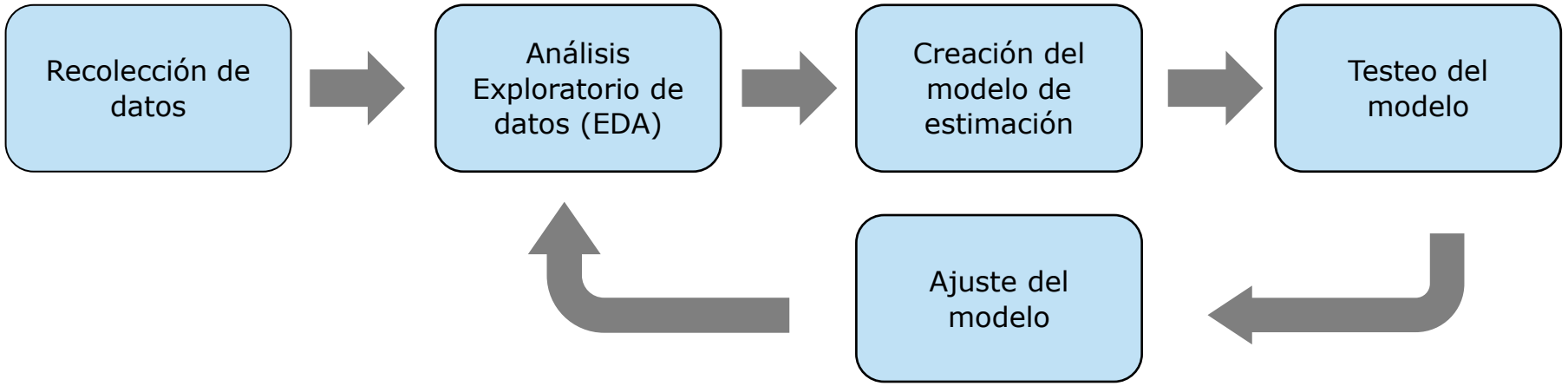
Debido a los requerimientos actuales de la compañía y la reciente creación del departamento de ciencia de datos, ha surgido la necesidad de construir una herramienta de cálculo para tasar potenciales propiedades de los clientes que puedan ingresar a nuestro sitio web; el objetivo entonces apunta en ayudar a nuestros clientes mejorando la experiencia y uso del sitio web que desemboque en más visitas y más propiedades que estén a cargo de nuestra compañía (inmobiliaria).

Esta herramienta interactiva debe ser capaz de estimar el costo de una propiedad según diversas características que el cliente pueda suministrar, por tal motivo, como parte de un plan piloto, el departamento de experiencia de cliente y ciencia de datos ha iniciado un análisis para construir un modelo de predicción que a priori permita estimar el costo de una vivienda y así determinar características claves, modelos a utilizar y posibles desempeños que generen una primera visión de solución a esta necesidad.

A continuación, se describirán ciertos elementos de análisis y metodológicos que desembocarán en un pliego de resultados que esclarezcan los próximos pasos a seguir para el asentamiento de este plan piloto.

# METODOLOGÍA

A nivel metodológico para este plan piloto se han considerado las siguientes etapas:




Cada etapa será explicada de forma resumida a continuación.

# METODOLOGÍA

A nivel metodológico para este plan piloto se han considerado las siguientes etapas:

| Etapa | Procedimiento |
|---|---|
| Recolección de datos | • Datos referenciales de Ames, Iowa (Fuente: Kaggle)<br>• Sets de datos de Entrenamiento y Pruebas<br>• Volumen de datos 79 variables (categóricas y numéricas) x 1460 filas |
| Análisis Exploratorio de datos (EDA) | • Análisis de características numéricas y categóricas<br>• Revisión de medidas estadísticas dentro de la data<br>• Imputación de valores nulos<br>• Tratamiento de Valores perdidos (outliers)<br>• Análisis de correlación de variables<br>• Selección de características Claves |

# METODOLOGÍA

A nivel metodológico para este plan piloto se han considerado las siguientes etapas:

| Etapa | Procedimiento |
|---|---|
| Creación del modelo de estimación | • Codificación de características (one-hot encoding)<br>• Split de data (data de entrenamiento y testeo del modelo)<br>• Escalamiento de los datos<br>• Modelado (Regresión Lineal Lasso)<br>• Obtención de métricas de desempeño del modelo |
| Testeo del modelo | • Ajuste de data de testeo en base a data de entrenamiento<br>• Codificación de características (one-hot encoding)<br>• Ejecución de predicciones<br>• Impresión de resultados y análisis de estos |
| Ajuste del modelo | • Ajuste de coeficientes del modelo<br>• Selección de más o menos características<br>• Validación de métodos para tratamientos de Outliers y N/A<br>• Definición de modelo final |

# RESULTADOS

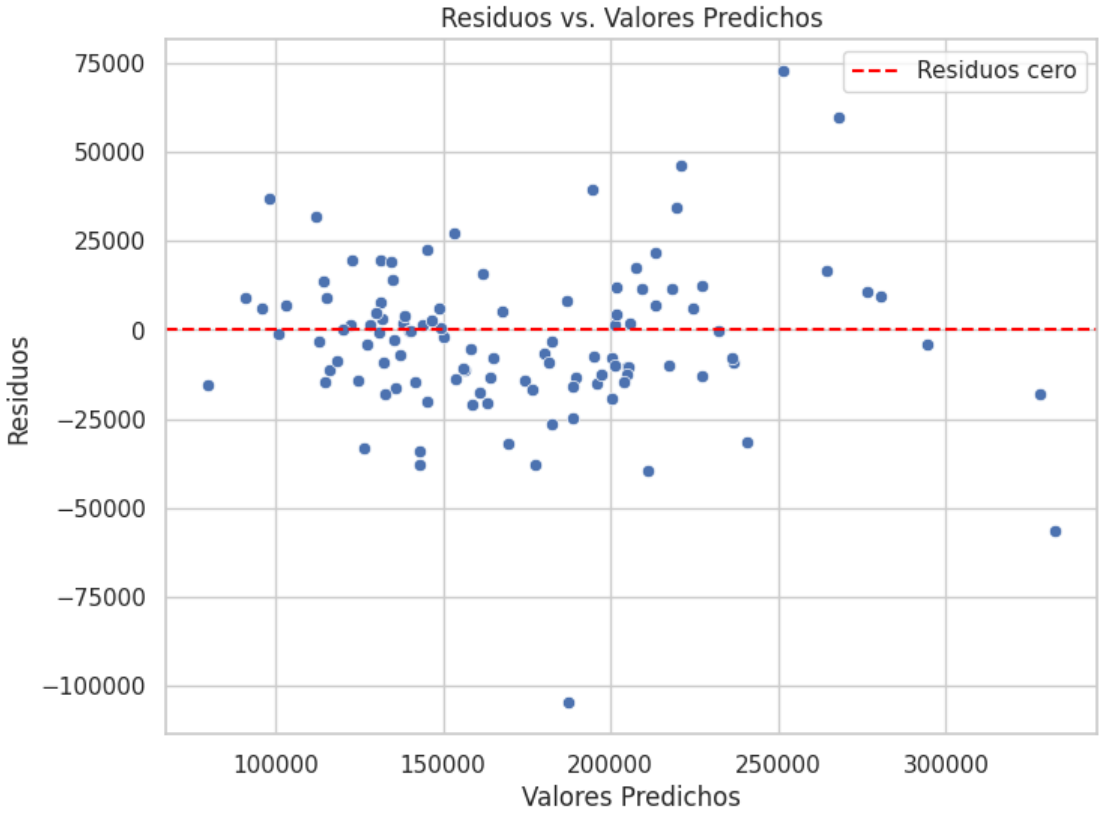
Al tratarse de una predicción cuyo valor objetivo es el precio de venta de una propiedad según ciertas características, concluimos:

- El modelo utilizado ha sido la regresión Lasso debido a la característica del problema, donde se requiere predecir una variable objetivo numérica como el precio de venta (SalePrice) de la propiedad. La ventaja de esta regresión es que permite la penalización de coeficientes de algunas variables a 0 que se traduce en una mejor selección de características, ventajoso frente a más de 70 características iniciales.
- La regresión Lasso es menos sensible al sobre ajuste por lo que la convierte en un buen modelo para estimar este tipo de escenarios.
- Las características numéricas con mayor incidencia en la predicción son: 'YearRemodAdd', 'TotalBsmtSF', 'FullBath', '1stFlrSF', 'TotRmsAbvGrd', 'GrLivArea', 'GarageCars', 'OverallQual', 'MasVnrArea', 'GarageYrBlt', 'YearBuilt', 'GarageArea'.
- Las características categóricas con mayor incidencia en la predicción son: 'MSZoning', 'Street', 'LotShape', 'LandContour', 'LotConfig', 'LandSlope', 'Neighborhood', 'Condition1', 'BldgType', 'HouseStyle', 'RoofStyle', 'Exterior1st', 'Exterior2nd', 'MasVnrType', 'ExterQual', 'ExterCond', 'Foundation', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'Heating', 'HeatingQC', 'CentralAir', 'Electrical', 'KitchenQual', 'Functional', 'FireplaceQu', 'GarageType', 'GarageFinish', 'GarageQual', 'GarageCond', 'PavedDrive', 'SaleType', 'SaleCondition'.

# RESULTADOS

Al tratarse de una predicción cuyo valor objetivo es el precio de venta de una propiedad según ciertas características, concluimos:

- El resultado de las predicciones es prometedor y como podemos ver en el siguiente gráfico, los residuos a 0 (coincidencias) se encuentran distribuidos uniformemente dentro de la nube de estimación, lo que conlleva a pensar que el modelo está representando adecuadamente la variabilidad de los datos.
- Respecto al indicador R^2 concluimos que existe un 83% de la variabilidad representada por el modelo en la predicción del precio de venta.



# RESULTADOS

Al tratarse de una predicción cuyo valor objetivo es el precio de venta de una propiedad según ciertas características, concluimos:

- Respecto al indicador RMSE concluimos que las predicciones del modelo tienen un error de alrededor de $21975 en comparación con los valores reales, que para en términos de un rango de SalePrice de ejemplo (Train) que va desde 34.900$ a 755.000$ representa un 5.8%.

| Fuente de datos | MSE | R^2 | RMSE |
|---|---|---|---|
| **Train (80%)** | 1.264730e+08 | 0.957033 | 11246.021868 |
| **Traing (20%)** | 4.829266e+08 | 0.836646 | 21975.589981 |

- Podemos agregar que se si bien se ha hecho un trabajo de feature engineering, tratando de identificar variables claves, tratamiento de outliers y valores nulos se sugiere una mayor cantidad de iteraciones con el objetivo de obtener mejores resultados frente a escenarios reales, y más iteraciones para reducir (en la medida de lo posible el número de variables).

- También se sugiere la implementación de modelos de ensamblado en bagging para contrastar los resultados previos y en caso de ser necesario llegar a un modelo más robusto.