Количественные

'LotFrontage' - Линейные футы улицы, подключенные к собственности. Слабо коррелирует с таргетом, возможно, удалять? Пропуски заполнялись средним по трейну и тесту;

'LotArea' - Размер участка в квадратных футах. Слабо коррелирует с таргетом, удалять?

'YearBuilt' - Первоначальная дата постройки (превратил в возраст дома, как дата продажи – дата постройки), дроп;

'YearRemodAdd' - Дата реконструкции (совпадает с датой строительства, если не было реконструкции или дополнений) (превратил в годы с ремонта), дроп;

'MasVnrArea' - Площадь облицовки каменной кладки в квадратных футах. Пропуски заполнялись средним по трейну и тесту. Кажется не очень информативной;

'BsmtFinSF1' – законченные кв. футы подвала? Пропуски на тесте заполнял средним по трейну и тесту;

'BsmtFinSF2' – законченные кв. футы подвала 2-го типа? Пропуски на тесте заполнял средним по трейну и тесту. Дроп, т. к. некоррелирует ни с чем и кажется бессмысленной.

'BsmtUnfSF' – незаконченные квадратные футы подвальной площади. Существенная отриц. Корреляция с BsmtFinSF1. Пропуски на тесте заполнял средним по трейну и тесту. Слабая корреляция с таргетом. Дроп.

'TotalBsmtSF' – общая площадь подвала в квадратных футах. Сильная полож. коррел. с 1stFlrSF. Пропуски на тесте заполнял средним по трейну и тесту.

'1stFlrSF' – площадь 1-го этажа в кв. футах. Дроп;

'2ndFlrSF' – площадь 2-го этажа в кв. футах. Дроп (без них качество лучше);

'LowQualFinSF' – Некачественная отделка квадратных футов (все этажи). Скаттерплот неинформативный. Переменная практически для всех домов равна нулю. Возможно, дроп?

'GrLivArea' – жилая площадь над уровнем земли;

'GarageYrBlt' – Вряд ли возраст гаража сильно сказывается на цене дома и, вероятно, что эта переменная должна как-то коррелировать с возрастом самого дома. Дроп.

'GarageArea' – площадь гаража в кв. футах.

'WoodDeckSF' – Площадь деревянной палубы в квадратных футах.

'OpenPorchSF' – Площадь открытой веранды в квадратных футах,

'EnclosedPorch' – Закрытая площадь веранды в квадратных футах,

'3SsnPorch' – площадь 3-х сезонной веранды в квадратных футах. Корреляция слабая. Скаттерплот просто облако. Дроп.

'ScreenPorch' – Площадь крыльца с экраном в квадратных футах. Дроп.,

'PoolArea' – дроп (маловариативный признак, у большинства нет бассейнов).

'MiscVal' – стоимость прочих функций, не включенных в другие категории. Маловариативна (мало домов с такими функциями, видимо), слабая корреляция. Дроп.

'YrSold' – год продажи (использовался для создания HouseAge, RepairAge), дроп;

Добавлены:

'HouseAge' – возраст дома на момент покупки

'RepairAge' – срок с последнего ремонта на момент покупки (дроп из-за высокой корреляции с 'HouseAge'

Категориальные (ординальные)

'BsmtFullBath' – количество полноценных ванных в подвале. Дроп.

'BsmtHalfBath' – количество половинных ванных в подвале. Дроп.

'FullBath' – количество полноценных ванных,

'HalfBath' – количество половинных ванных. Дроп.

'BedroomAbvGr' – спален. Дроп.

'KitchenAbvGr' – количество кухонь над землей,

'TotRmsAbvGrd' – количество комнат над землей,

'Fireplaces' – количество каминов,

'GarageCars' – вместимость гаража (кол-во машин)

Категориальные (номинальные)

'MSZoning' – тип зоны (индустриал., коммерческая, жилая с высок. Плотностью). Пропуски в тесте заполнил модой по тесту.

'Street' – тип подъездной дороги (гравий или мощеная). Почти везде (99%) мощеная. Дроп

'Alley' – тип подъездной аллеи к собственности (гравий или мощеная). Слабовариативная, без нее качество лучше. Дроп

'LotShape',

'LandContour',

'Utilities' – доступные удобства (газ, электрич., вода). Заполнил пропуск в тесте модой. Признак почти константный. Дроп.

'LotConfig',

'LandSlope' – уклон собственности (земли, видимо) (легкий, значительный, сильный);

'Neighborhood',

'Condition1' – наличие поблизости от дома каких-то особых условий (рядом с ж/д, парком, крупной дорогой);

'Condition2' – почти константная. Дроп

'BldgType',

'HouseStyle',

'RoofStyle',

'RoofMatl' – материал крыши,

'Exterior1st' – Наружное покрытие дома,

'Exterior2nd' – Наружное покрытие дома (если больше 1 материала). Этот и предыд. Заполнил в тесте модой

'MasVnrType' - тип облицовки каменной кладки. Пропуски заполнялись модой и в трейне, и в тесте.

'ExterQual',

'ExterCond',

'Foundation',

'BsmtQual',

'BsmtCond',

'BsmtExposure',

'BsmtFinType1',

'BsmtFinType2',

'Heating',

'HeatingQC',

'CentralAir',

'Electrical' – тип электрич. Системы. Пропуск в трейне заменил модой.

'KitchenQual' – качество кухни. Пропуск в тесте модой;

'Functional' – пригодность к жизни (степень поврежденности). Пропуски на тесте модой;

'FireplaceQu',

'GarageType',

'GarageFinish',

'GarageQual',

'GarageCond',

'PavedDrive' – мощеная ли подъездная дорожка (да/частично/гравий или грунт);

'PoolQC' – качество бассейна. Почти все пропуски, кроме нескольких строчек. Заполнил словом «No», но, скорее всего, дроп.

'Fence' – качество забора. Бокс-плот не особо информативный, по смыслу кажется не особо влияющей. Дроп

'MiscFeature' – прочие функции, не включенные в другие категории. Маловариативна (мало домов с такими функциями, видимо). Дроп.

'SaleType' – тип сделки. Пропуск на тесте модой;

'SaleCondition' – условие сделки (обычная, присоединение прилегающей земли, сделка между родственниками и т. д.),

'MSSubClass' – тип жилья (одноэтажный, 2-этажный, дюплекс и т. д.);

'OverallQual' – качество материалов и отделки дома;

'OverallCond' – общее качество дома;

'MoSold' – месяц продажи

Целевая переменная

SalePrice – цена продажи (исходный таргет);

log_SalePrice – логарифм цены продажи.