In [1]:

```
import numpy as np
import pandas as pd
```

In [2]:

```
x_train = pd.read_csv('x_train.csv', sep = ',',index_col=0)
y_train = pd.read_csv('y_train.csv', sep = ',',names=['price'],index_col=0)
x_test = pd.read_csv('x_test.csv', sep = ',',index_col=0)
y_test = pd.read_csv('y_test.csv', sep = ',',names=['price'],index_col=0)
```

In [3]:

```
x_train.head()
```

Out[3]:

	train_id	name	item_condition_id	category_name	brar
496798	496798	Tan cardigan, size medium	2	Women/Sweaters/Cardigan	Max
1315605	1315605	Picture frame	1	Home/Home Décor/Photo Albums & Frames	NaN
1104183	1104183	ibloom english bread	2	Vintage & Collectibles/Antique/Collectibles	NaN
424705	424705	Terry's chocolate oranges	1	Home/Kitchen & Dining/Coffee & Tea Accessories	NaN
145825	145825	Toms Wedge Heels Size 6	3	Women/Shoes/Sandals	TOM

In [4]:

```
y_train.head()
```

Out[4]:

	price
496798	14.0
1315605	7.0
1104183	30.0
424705	24.0
145825	17.0

Item Condition

In [5]:

```
x_train['item_condition_id'].value_counts()
```

Out[5]:

```
1    313988
3    211724
2    183936
4     15653
5      1140
Name: item_condition_id, dtype: int64
```

In [6]:

```
x_train['item_condition_id'].isnull().sum()
```

Out[6]:

```
0
```

In [7]:

```
y_train.head()
```

Out[7]:

	price
496798	14.0
1315605	7.0
1104183	30.0
424705	24.0
145825	17.0

In [8]:

```
# response coding on item_condition ## TRAIN DATA ##

temp = (x_train['item_condition_id']==1).sum()
x = (y_train.loc[x_train['item_condition_id']==1].sum(axis = 0)/temp.sum())[0]
z = x_train['item_condition_id'] == 1
x_train.item_condition_id.loc[z] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)
```

In [9]:

```
# response coding on item_condition ## TEST DATA ##
temp = (x_test['item_condition_id']==1).sum()
x = (y_test.loc[x_test['item_condition_id']==1].sum(axis = 0)/temp.sum())[0]
z = x_test['item_condition_id'] == 1
x_test.item_condition_id.loc[z] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)
```

In [10]:

```
# train
temp = (x_train['item_condition_id']==2).sum()
x = (y_train.loc[x_train['item_condition_id']==2].sum(axis = 0)/(x_train['item_c
ondition_id']==2).sum())[0]
z = x_train['item_condition_id'] == 2
x_train.item_condition_id.loc[z] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)
```

In [11]:

```
# test
temp = (x_test['item_condition_id']==2).sum()
x = (y_test.loc[x_test['item_condition_id']==2].sum(axis = 0)/(x_test['item_cond
ition_id']==2).sum())[0]
z = x_test['item_condition_id'] == 2
x_test.item_condition_id.loc[z] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)
```

In [12]:

```
#train
temp = (x_train['item_condition_id']==3).sum()
x = (y_train.loc[x_train['item_condition_id']==3].sum(axis = 0)/(x_train['item_c
ondition_id']==3).sum())[0]
z = x_train['item_condition_id'] == 3
x_train.item_condition_id.loc[z] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)
```

In [13]:

```
#test
temp = (x_test['item_condition_id']==3).sum()
x = (y_test.loc[x_test['item_condition_id']==3].sum(axis = 0)/(x_test['item_cond
ition_id']==3).sum())[0]
z = x_test['item_condition_id'] == 3
x_test.item_condition_id.loc[z] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)
```

In [14]:

```
#train
temp = (x_train['item_condition_id']==4).sum()
x = (y_train.loc[x_train['item_condition_id']==4].sum(axis = 0)/(x_train['item_c
ondition_id']==4).sum())[0]
z = x_train['item_condition_id'] == 4
x_train.item_condition_id.loc[z] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)
```

In [15]:

```
#test
temp = (x_test['item_condition_id']==4).sum()
x = (y_test.loc[x_test['item_condition_id']==4].sum(axis = 0)/(x_test['item_cond
ition_id']==4).sum())[0]
z = x_test['item_condition_id'] == 4
x_test.item_condition_id.loc[z] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)
```

In [16]:

```
#train
temp = (x_train['item_condition_id']==2).sum()
x = (y_train.loc[x_train['item_condition_id']==2].sum(axis = 0)/(x_train['item_c
ondition_id']==2).sum())[0]
z = x_train['item_condition_id'] == 2
x_train.item_condition_id.loc[z] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)
```

In [17]:

```
#test
temp = (x_test['item_condition_id']==2).sum()
x = (y_test.loc[x_test['item_condition_id']==2].sum(axis = 0)/(x_test['item_cond
ition_id']==2).sum())[0]
z = x_test['item_condition_id'] == 2
x_test.item_condition_id.loc[z] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)
```

In [18]:

```
#train
temp = (x_train['item_condition_id']==5).sum()
x = (y_train.loc[x_train['item_condition_id']==5].sum(axis = 0)/(x_train['item_c
ondition_id']==5).sum())[0]
z = x_train['item_condition_id'] == 5
x_train.item_condition_id.loc[z] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)
```

In [19]:

```
#test
temp = (x_test['item_condition_id']==5).sum()
x = (y_test.loc[x_test['item_condition_id']==5].sum(axis = 0)/(x_test['item_cond
ition_id']==5).sum())[0]
z = x_test['item_condition_id'] == 5
x_test.item_condition_id.loc[z] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)
```

In [ ]:

```

```

Category

In [20]:

```
def split_cat(text):
    try: return text.split("/")
    except: return ("No Label", "No Label", "No Label")
```

In [21]:

```
#train
# splitting the raw category into main and sub sub categories
x_train['main_cat'], x_train['subcat_1'], x_train['subcat_2'] = \
zip(*x_train['category_name'].apply(lambda x: split_cat(x)))
# sanity checking the train for new categories
x_train.head()
```

Out[21]:

	train_id	name	item_condition_id	category_name	brar
496798	496798	Tan cardigan, size medium	27.514021	Women/Sweaters/Cardigan	Max
1315605	1315605	Picture frame	26.399296	Home/Home Décor/Photo Albums & Frames	NaN
1104183	1104183	ibloom english bread	27.514021	Vintage & Collectibles/Antique/Collectibles	NaN
424705	424705	Terry's chocolate oranges	26.399296	Home/Kitchen & Dining/Coffee & Tea Accessories	NaN
145825	145825	Toms Wedge Heels Size 6	26.458888	Women/Shoes/Sandals	TOM

In [22]:

```
#test
# splitting the raw category into main and sub sub categories
x_test['main_cat'], x_test['subcat_1'], x_test['subcat_2'] = \
zip(*x_test['category_name'].apply(lambda x: split_cat(x)))
# sanity checking the train for new categories
x_test.head()
```

Out[22]:

	train_id	name	item_condition_id	category_name	brand_name	ship
777341	777341	F/ship 4 Totoro Washi + 1 pen	26.530201	Handmade/Paper Goods/Stationery	NaN	1
1463629	1463629	UCLA Men's Bundle + Shorts	26.530201	Women/Other/Other	Adidas	1
350669	350669	Listing for lol	26.530201	Beauty/Makeup/Lips	NaN	1
310222	310222	25 pcs kawaii sticker flakes	26.530201	Kids/Toys/Arts & Crafts	NaN	1
759257	759257	Chanel Mini Lipgloss Set	27.685444	Beauty/Makeup/Lips	Chanel	1

In [23]:

```
print("Train data")
print(x_train['main_cat'].isnull().sum())
print(x_train['subcat_1'].isnull().sum())
print(x_train['subcat_1'].isnull().sum())

print("Test data")
print(x_test['main_cat'].isnull().sum())
print(x_test['subcat_1'].isnull().sum())
print(x_test['subcat_1'].isnull().sum())
```

```
Train data
0
0
0
Test data
0
0
0
```

In [ ]:

```
```

In [24]:

```
#train
x_train['main_cat'].nunique()
```

Out[24]:

```
11
```

In [25]:

```
#test
x_test['main_cat'].nunique()
```

Out[25]:

```
11
```

In [26]:

```
#train
x_train['subcat_1'].nunique()
```

Out[26]:

```
114
```

In [27]:

```
#test
x_test['subcat_1'].nunique()
```

Out[27]:

```
114
```

In [28]:

```
#train
x_train['subcat_2'].nunique()
```

Out[28]:

```
840
```

In [29]:

```
#test
x_test['subcat_2'].nunique()
```

Out[29]:

```
807
```

In [30]:

```
%%time
#train
lk = dict()
for cat in x_train['main_cat'].unique():
    try:
        if lk[cat]:
            z = x_train['main_cat']==cat
            x_train.main_cat.loc[z] = lk[cat]
    except:
        temp =(x_train['main_cat']==cat).sum()
        x = (y_train.loc[x_train['main_cat']==cat].sum(axis = 0)/temp)[0]
        z = x_train['main_cat']==cat
        x_train.main_cat.loc[z] = x
        lk[cat] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)

CPU times: user 3.78 s, sys: 111 ms, total: 3.89 s
Wall time: 3.88 s
```

In [31]:

```
%%time
#test
lkk = dict()
for cat in x_test['main_cat'].unique():
    try:
        if lkk[cat]:
            z = x_test['main_cat']==cat
            x_test.main_cat.loc[z] = lkk[cat]
    except:
        temp =(x_test['main_cat']==cat).sum()
        x = (y_test.loc[x_test['main_cat']==cat].sum(axis = 0)/temp)[0]
        z = x_test['main_cat']==cat
        x_test.main_cat.loc[z] = x
        lkk[cat] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)

CPU times: user 2.43 s, sys: 51.9 ms, total: 2.49 s
Wall time: 2.48 s
```

In [ ]:

```

```

In [ ]:

```
```

In [32]:

```
%%time
#train
lk1 = dict()
for cat in x_train['subcat_1'].unique():
    try:
        if lk1[cat]:
            z = x_train['subcat_1']==cat
            x_train.subcat1.loc[z] = lk1[cat]
    except:
        temp =(x_train['subcat_1']==cat).sum()
        x = (y_train.loc[x_train['subcat_1']==cat].sum(axis = 0)/temp)[0]
        z = x_train['subcat_1']==cat
        x_train.subcat_1.loc[z] = x
        lk1[cat] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)

CPU times: user 29.4 s, sys: 17.9 ms, total: 29.4 s
Wall time: 29.4 s
```

In [33]:

```
%%time
#test
lkk1 = dict()
for cat in x_test['subcat_1'].unique():
    try:
        if lkk1[cat]:
            z = x_test['subcat_1']==cat
            x_test.subcat1.loc[z] = lkk1[cat]
    except:
        temp =(x_test['subcat_1']==cat).sum()
        x = (y_test.loc[x_test['subcat_1']==cat].sum(axis = 0)/temp)[0]
        z = x_test['subcat_1']==cat
        x_test.subcat_1.loc[z] = x
        lkk1[cat] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)

CPU times: user 21.3 s, sys: 19.7 ms, total: 21.3 s
Wall time: 21.3 s
```

In [34]:

```
x_train.head()
```

Out[34]:

	train_id	name	item_condition_id	category_name	brand_
496798	496798	Tan cardigan, size medium	27.514021	Women/Sweaters/Cardigan	Max St
1315605	1315605	Picture frame	26.399296	Home/Home Décor/Photo Albums & Frames	NaN
1104183	1104183	ibloom english bread	27.514021	Vintage & Collectibles/Antique/Collectibles	NaN
424705	424705	Terry's chocolate oranges	26.399296	Home/Kitchen & Dining/Coffee & Tea Accessories	NaN
145825	145825	Toms Wedge Heels Size 6	26.458888	Women/Shoes/Sandals	TOMS

In [35]:

```
"""%%time
for cat in x_train['subcat_1'].unique():
    temp =(x_train['subcat_1']==cat).sum()
    x = (y_train.loc[x_train['subcat_1']==cat].sum(axis = 0)/(x_train['subcat_
1']==cat).sum())[0]
    z = x_train['subcat_1']==cat
    x_train.subcat_1.loc[z] = x"""
```

Out[35]:

```
"%%time\nfor cat in x_train['subcat_1'].unique():\n    temp =(x_trai
n['subcat_1']==cat).sum()\n    x = (y_train.loc[x_train['subcat_1']=
=cat].sum(axis = 0)/(x_train['subcat_1']==cat).sum())[0]\n    z = x_
train['subcat_1']==cat\n    x_train.subcat_1.loc[z] = x"
```

In [36]:

```
%%time
#train
lk2 = dict()
for cat in x_train['subcat_2'].unique():
    try:
        if lk2[cat]:
            z = x_train['subcat_2']==cat
            x_train.subcat_2.loc[z] = lk2[cat]
    except:
        temp =(x_train['subcat_2']==cat).sum()
        x = (y_train.loc[x_train['subcat_2']==cat].sum(axis = 0)/temp)[0]
        z = x_train['subcat_2']==cat
        x_train.subcat_2.loc[z] = x
        lk2[cat] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)

CPU times: user 3min 11s, sys: 138 ms, total: 3min 11s
Wall time: 3min 11s
```

In [37]:

```
%%time
#test
lkk2 = dict()
for cat in x_test['subcat_2'].unique():
    try:
        if lkk2[cat]:
            z = x_test['subcat_2']==cat
            x_test.subcat_2.loc[z] = lkk2[cat]
    except:
        temp =(x_test['subcat_2']==cat).sum()
        x = (y_test.loc[x_test['subcat_2']==cat].sum(axis = 0)/temp)[0]
        z = x_test['subcat_2']==cat
        x_test.subcat_2.loc[z] = x
        lkk2[cat] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)

CPU times: user 2min 17s, sys: 104 ms, total: 2min 17s
Wall time: 2min 17s
```

In [38]:

```
x_train.head()
```

Out[38]:

	train_id	name	item_condition_id	category_name	bra
496798	496798	Tan cardigan, size medium	27.514021	Women/Sweaters/Cardigan	Max
1315605	1315605	Picture frame	26.399296	Home/Home Décor/Photo Albums & Frames	NaN
1104183	1104183	ibloom english bread	27.514021	Vintage & Collectibles/Antique/Collectibles	NaN
424705	424705	Terry's chocolate oranges	26.399296	Home/Kitchen & Dining/Coffee & Tea Accessories	NaN
145825	145825	Toms Wedge Heels Size 6	26.458888	Women/Shoes/Sandals	TOM

In [39]:

```
"""%%time
for cat in x_train['subcat_2'].unique():
    temp =(x_train['subcat_2']==cat).sum()
    x = (y_train.loc[x_train['subcat_2']==cat].sum(axis = 0)/(x_train['subcat_
2']==cat).sum())[0]
    z = x_train['subcat_2']==cat
    x_train.subcat_2.loc[z] = x"""
```

Out[39]:

```
"%%time\nfor cat in x_train['subcat_2'].unique():\n    temp =(x_trai
n['subcat_2']==cat).sum()\n    x = (y_train.loc[x_train['subcat_2']=
=cat].sum(axis = 0)/(x_train['subcat_2']==cat).sum())[0]\n    z = x_
train['subcat_2']==cat\n    x_train.subcat_2.loc[z] = x"
```

In [ ]:

```

```

Brand Name

In [40]:

```
x_train['brand_name'].isnull().sum()
```

Out[40]:

```
310239
```

In [41]:

```
x_train['isBrandNull'] = x_train['brand_name'].fillna(1)
```

In [42]:

```
x_train['isBrandNull'] = x_train.apply(
    lambda row: 0 if pd.notnull(row['isBrandNull']) and (row['isBrandNull'] != 1
) else row['isBrandNull'],
    axis=1
)
```

In [43]:

```
x_train['isBrandNull'].head()
```

Out[43]:

```
496798     0
1315605    1
1104183    1
424705     1
145825     0
Name: isBrandNull, dtype: int64
```

In [44]:

```
#train
x_train['brand_name'].nunique()
```

Out[44]:

```
3997
```

In [45]:

```
#test
x_test['brand_name'].nunique()
```

Out[45]:

```
3367
```

In [46]:

```
x_test['brand_name'].isnull().sum()
```

Out[46]:

```
189782
```

In [47]:

```
x_test['isBrandNull'] = 0
```

In [48]:

```
x_test['isBrandNull'].head()
```

Out[48]:

```
777341     0
1463629    0
350669     0
310222     0
759257     0
Name: isBrandNull, dtype: int64
```

In [ ]:

```

```

In [49]:

```
%%time
#train
look = dict()
counter = 0
for cat in x_train['brand_name'].unique():
    counter+=1
    if(counter == 500):
        print("500 iterations completed")
        counter = 0
    try:
        if look[cat]:
            z = x_train['brand_name']==cat
            x_train.brand_name.loc[z] = look[cat]
    except:
        temp =(x_train['brand_name']==cat).sum()
        x = (y_train.loc[x_train['brand_name']==cat].sum(axis = 0)/temp)[0]
        z = x_train['brand_name']==cat
        x_train.brand_name.loc[z] = x
        look[cat] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)

500 iterations completed
500 iterations completed
500 iterations completed
500 iterations completed
500 iterations completed
500 iterations completed
500 iterations completed
CPU times: user 11min 52s, sys: 464 ms, total: 11min 53s
Wall time: 11min 53s
```

In [50]:

```
x_train.head()
```

Out[50]:

	train_id	name	item_condition_id	category_name	brand_
496798	496798	Tan cardigan, size medium	27.514021	Women/Sweaters/Cardigan	15.638
1315605	1315605	Picture frame	26.399296	Home/Home Décor/Photo Albums & Frames	NaN
1104183	1104183	ibloom english bread	27.514021	Vintage & Collectibles/Antique/Collectibles	NaN
424705	424705	Terry's chocolate oranges	26.399296	Home/Kitchen & Dining/Coffee & Tea Accessories	NaN
145825	145825	Toms Wedge Heels Size 6	26.458888	Women/Shoes/Sandals	23.284

In [51]:

```
%%time
#test
look1 = dict()
counter = 0
for cat in x_test['brand_name'].unique():
    counter = counter + 1
    if(counter == 500):
        print("500 iterations completed")
        counter = 0
    try:
        if look1[cat]:
            z = x_test['brand_name']==cat
            x_test.brand_name.loc[z] = look1[cat]
    except:
        temp =(x_test['brand_name']==cat).sum()
        x = (y_test.loc[x_test['brand_name']==cat].sum(axis = 0)/temp)[0]
        z = x_test['brand_name']==cat
        x_test.brand_name.loc[z] = x
        look1[cat] = x
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/pandas/core/i
ndexing.py:194: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)

500 iterations completed
500 iterations completed
500 iterations completed
500 iterations completed
500 iterations completed
500 iterations completed
CPU times: user 7min 39s, sys: 424 ms, total: 7min 39s
Wall time: 7min 39s
```

In [ ]:

```

```

In [52]:

```
"""%%time
#test
look1 = dict()
counter = 0
for cat in x_test['brand_name'].unique():
    counter+=1
    if(counter == 500):
        print("500 iterations completed")
        counter = 0
    try:
        if look1[cat]:
            z = x_test['brand_name']==cat
            x_test.brand_name.loc[z] = look1[cat]
    except:
        temp =(x_test['brand_name']==cat).sum()
        x = (y_test.loc[x_test['brand_name']==cat].sum(axis = 0)/temp)[0]
        z = x_test['brand_name']==cat
        x_test.brand_name.loc[z] = x
        look1[cat] = x"""
```

Out[52]:

```
'%%time\n#test\nlook1 = dict()\ncounter = 0\nfor cat in x_test[\'bra
nd_name\'].unique():\n    counter+=1\n    if(counter == 500):\n
   print("500 iterations completed")\n        counter = 0\n    tr
y:\n        if look1[cat]:\n            z = x_test[\'brand_name\']==
cat\n            x_test.brand_name.loc[z] = look1[cat]\n    except:
\n        temp =(x_test[\'brand_name\']==cat).sum()\n        x = (y_
test.loc[x_test[\'brand_name\']==cat].sum(axis = 0)/temp)[0]\n
  z = x_test[\'brand_name\']==cat\n        x_test.brand_name.loc[z]
= x\n        look1[cat] = x'
```

In [53]:

```
#train
mean = (y_train.mean())[0]
```

In [54]:

```
#test
meanT = (y_test.mean())[0]
```

In [55]:

```
x_train['brand_name'] = x_train['brand_name'].fillna(mean)
```

In [56]:

```
x_test['brand_name'] = x_test['brand_name'].fillna(meanT)
```

In [57]:

```
x_train.head()
```

Out[57]:

	train_id	name	item_condition_id	category_name	brand_
496798	496798	Tan cardigan, size medium	27.514021	Women/Sweaters/Cardigan	15.638
1315605	1315605	Picture frame	26.399296	Home/Home Décor/Photo Albums & Frames	26.660
1104183	1104183	ibloom english bread	27.514021	Vintage & Collectibles/Antique/Collectibles	26.660
424705	424705	Terry's chocolate oranges	26.399296	Home/Kitchen & Dining/Coffee & Tea Accessories	26.660
145825	145825	Toms Wedge Heels Size 6	26.458888	Women/Shoes/Sandals	23.284

In [ ]:

```
```

In [58]:

```
from sklearn.feature_extraction import stop_words
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.corpus import stopwords
import re
```

In [59]:

```
stop = set(stopwords.words('english'))
def tokenizeW(text):
    """
    sent_tokenize(): segment text into sentences
    word_tokenize(): break sentences into words
    """
    try:
        regex = re.compile('[^A-Za-z0-9]+')
        text = regex.sub(" ", text) # remove punctuation

        tokens_ = [word_tokenize(s) for s in sent_tokenize(text)]
        tokens = []
        for token_by_sent in tokens_:
            tokens += token_by_sent
        tokens = list(filter(lambda t: t.lower() not in stop, tokens))
        filtered_tokens = [w for w in tokens if re.search('[a-zA-Z]', w)]
        filtered_tokens = [w.lower() for w in filtered_tokens]

        return filtered_tokens

    except TypeError as e: print(text,e)
```

In [60]:

```
x_train['item_description'].isnull().sum()
```

Out[60]:

```
1
```

In [61]:

```
y_train.shape
```

Out[61]:

```
(726441, 1)
```

In [62]:

```
yy = y_train.loc[x_train['item_description'].isnull()].copy()
```

In [63]:

```
ind = yy.index.values
y_train.drop(ind,inplace = True)
```

In [64]:

```
y_train.shape
```

Out[64]:

```
(726440, 1)
```

In [65]:

```
#x_train.dropna( how='any',inplace = True)
x_train.dropna(axis=0, subset=['item_description'], thresh=1,inplace = True)
```

In [66]:

```
x_train.shape
```

Out[66]:

(726440, 11)

In [67]:

```
x_train['item_description'].isnull().sum()
```

Out[67]:

0

In [68]:

```
x_test.head()
```

Out[68]:

	train_id	name	item_condition_id	category_name	brand_name	sh
777341	777341	F/ship 4 Totoro Washi + 1 pen	26.530201	Handmade/Paper Goods/Stationery	26.821436	1
1463629	1463629	UCLA Men's Bundle + Shorts	26.530201	Women/Other/Other	43.858812	1
350669	350669	Listing for lol	26.530201	Beauty/Makeup/Lips	26.821436	1
310222	310222	25 pcs kawaii sticker flakes	26.530201	Kids/Toys/Arts & Crafts	26.821436	1
759257	759257	Chanel Mini Lipgloss Set	27.685444	Beauty/Makeup/Lips	81.530612	1

In [69]:

```
#test
x_test['item_description'].isnull().sum()
```

Out[69]:

2

In [70]:

```
x_test.loc[x_test['item_description'].isnull()]
```

Out[70]:

	train_id	name	item_condition_id	category_name	brand_name	s
1264242	1264242	For Bianca	26.653718	Women/Women's Accessories/Scarves & Wraps	26.821436	1
511535	511535	Shoes for Michelle	24.346695	Kids/Girls 0-24 Mos/Shoes	26.821436	0

In [71]:

```
y_test.shape
```

Out[71]:

(444761, 1)

In [72]:

```
yy1 = y_test.loc[x_test['item_description'].isnull()].copy()
```

In [73]:

```
ind1 = yy1.index.values
y_test.drop(ind1,inplace = True)
```

In [74]:

```
y_test.shape
```

Out[74]:

(444759, 1)

In [75]:

```
x_test.dropna(axis=0, subset=['item_description'], thresh=1,inplace = True)
```

In [76]:

```
x_test.shape
```

Out[76]:

(444759, 11)

In [77]:

```
x_train.shape
```

Out[77]:

(726440, 11)

In [78]:

```
#train
tok_train = x_train['item_description'].map(tokenizeW).tolist()
```

In [79]:

```
#test
tok_test = x_test['item_description'].map(tokenizeW).tolist()
```

In [80]:

```
import gensim
from gensim.models import KeyedVectors
```

In [81]:

```
#train
w2v_model=gensim.models.Word2Vec(tok_train,min_count=1,size=75, workers=8)
```

In [82]:

```
#test
w2v_modelT=gensim.models.Word2Vec(tok_test,min_count=1,size=75, workers=8)
```

In [83]:

```
#train
w2v_model.save('w2v_model')
```

In [84]:

```
#test
w2v_modelT.save('w2v_modelT')
```

In [85]:

```
#train
model = gensim.models.Word2Vec.load('w2v_model')
```

In [86]:

```
#test
modelT = gensim.models.Word2Vec.load('w2v_modelT')
```

In [87]:

```
model.wv['pink']
```

Out[87]:

```
array([-1.16716909, -1.60554433,  1.21553171, -0.64954811, -0.397989
66,
       -1.42583334,  1.44479489,  1.32937968, -1.2875973 , -0.793595
55,
       -0.85141361,  0.75281852,  3.93656993, -0.94617325,  4.007638
93,
        2.26004934,  2.16077733, -2.06566548,  2.05004668, -1.386961
34,
        0.42800003,  1.9215728 ,  0.34223819, -0.54955274, -1.25132
  ,
       -0.19523257,  0.52542567,  0.76414943, -1.82838774,  1.999951
72,
        4.16405106,  0.51841706, -2.27938986,  0.37259546,  1.606081
01,
       -0.0500126 , -1.62154949,  0.36300898, -3.3543036 ,  2.416861
06,
        1.96871865, -1.12328875,  1.23098493, -3.05123401,  0.636088
61,
       -0.03036699,  0.08411524,  1.82481313, -0.4512468 , -1.969470
26,
       -0.75292915, -0.76885712, -4.70269823, -2.40906525,  0.501678
41,
        2.70678568, -1.48815179, -3.2877717 , -1.18626988,  0.507767
62,
       -0.5169493 , -3.60145426, -0.67325675,  2.12545586, -1.560465
69,
        2.04489684, -1.43300343, -0.03460142,  0.87221515, -0.643180
49,
        2.57997847, -1.90737522, -3.14051461,  3.28603125,  0.616720
14], dtype=float32)
```

In [88]:

```
words = list(w2v_model.wv.vocab)
print(len(words))
```

```
106148
```

In [89]:

```
words[:5]
```

Out[89]:

```
['pr5', '54diapers', 'hoodless', '7for', 'coolbourne']
```

In [90]:

```
w2v_model.wv['family']
```

Out[90]:

```
array([-0.54600251, -0.82250625, -0.88786626, -0.25276649, -1.140770
91,
        0.05533401,  1.57056832,  1.46315598, -1.45123351, -1.016870
86,
        1.12046039,  0.11105274, -2.37318325,  0.78750414,  0.729638
16,
       -1.57642436, -1.46863961,  3.49937296,  0.46746773,  0.449169
28,
       -1.90448427,  0.11089415,  3.25538421, -0.50899476, -1.259611
25,
       -2.17369938,  2.8740406 , -0.47765929,  0.46993813,  1.138625
62,
       -1.88689435, -1.70431972,  2.45847201, -0.54637337, -1.800024
63,
        0.05418036, -0.72124362,  2.32884693,  2.26861334, -2.440502
88,
        0.46336713,  0.70196098, -1.11431694,  1.30717456,  1.150980
71,
        0.44821885,  0.20322214, -2.6149044 , -0.57657474,  1.483253
24,
       -1.36725903,  0.74839193, -0.45754001,  2.18983054, -0.765794
28,
       -0.98820293, -1.40949965, -0.94681472, -0.92394596, -1.189633
49,
        3.128227  , -1.8313508 ,  1.30024338,  0.29507527,  0.710727
93,
       -0.38475645, -0.08153762,  1.41530776, -2.78519797, -1.245885
97,
       -0.13181928, -0.74068946,  1.6219548 , -1.62269378, -1.977474
69], dtype=float32)
```

In [91]:

```
X = model[model.wv.vocab]
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/ipykernel_lau
ncher.py:1: DeprecationWarning: Call to deprecated `__getitem__` (Me
thod will be removed in 4.0.0, use self.wv.__getitem__() instead).
  """Entry point for launching an IPython kernel.
```

In [ ]:

```

```

In [92]:

```
#train
lis =()
counter = 0
for sent in tok_train:
    x = 0
    counter = counter + 1
    if counter == 60000:
        print(counter)
        counter = 0
    for w in sent:
        x = x +  model.wv[w].sum()
    lis = np.append(lis,x)
```

```
60000
60000
60000
60000
60000
60000
60000
60000
60000
60000
60000
60000
```

In [93]:

```
#test
lisT =()
counter = 0
for sent in tok_test:
    x = 0
    counter = counter + 1
    if counter == 60000:
        print(counter)
        counter = 0
    for w in sent:
        x = x +  modelT.wv[w].sum()
    lisT = np.append(lisT,x)
```

```
60000
60000
60000
60000
60000
60000
60000
```

In [94]:

```
#train
se = pd.Series(lis)
```

In [95]:

```
#test
seT = pd.Series(lisT)
```

In [96]:

```
#train
x_train['descp_num_w2v'] = se.values
```

In [97]:

```
#test
x_test['descp_num_w2v'] = seT.values
```

In [98]:

```
x_train.head()
```

Out[98]:

	train_id	name	item_condition_id	category_name	brar
496798	496798	Tan cardigan, size medium	27.514021	Women/Sweaters/Cardigan	15.6
1315605	1315605	Picture frame	26.399296	Home/Home Décor/Photo Albums & Frames	26.6
1104183	1104183	ibloom english bread	27.514021	Vintage & Collectibles/Antique/Collectibles	26.6
424705	424705	Terry's chocolate oranges	26.399296	Home/Kitchen & Dining/Coffee & Tea Accessories	26.6
145825	145825	Toms Wedge Heels Size 6	26.458888	Women/Shoes/Sandals	23.2

In [99]:

```
#import pickle
```

In [100]:

```
#model = KeyedVectors.load_word2vec_format('GoogleNews-vectors-negative300.bin',
 binary=True)
```

In [ ]:

```
```

Name

In [ ]:

```
```

In [101]:

```
#train
x_train['name'].isnull().sum()
```

Out[101]:

```
0
```

In [102]:

```
#test
x_test['name'].isnull().sum()
```

Out[102]:

```
0
```

In [103]:

```
#train
tok_name = x_train['name'].map(tokenizeW).tolist()
```

In [104]:

```
#test
tok_nameT = x_test['name'].map(tokenizeW).tolist()
```

In [105]:

```
#train
w2v_model1=gensim.models.Word2Vec(tok_name,min_count=1,size=75, workers=8)
```

In [106]:

```
#test
w2v_model1T=gensim.models.Word2Vec(tok_nameT,min_count=1,size=75, workers=8)
```

In [107]:

```
#train
w2v_model1.save('w2v_model1')
```

In [108]:

```
#test
w2v_model1T.save('w2v_model1T')
```

In [109]:

```
#train
model1 = gensim.models.Word2Vec.load('w2v_model1')
```

In [110]:

```
#test
model1T = gensim.models.Word2Vec.load('w2v_model1T')
```

In [ ]:

```

```

In [111]:

```
#train
lis1 =()
counter = 0
for sent in tok_name:
    x = 0
    counter = counter + 1
    if counter == 60000:
        print(counter)
        counter = 0
    for w in sent:
        x = x +  model1.wv[w].sum()
    lis1 = np.append(lis1,x)
```

```
60000
60000
60000
60000
60000
60000
60000
60000
60000
60000
60000
60000
60000
```

In [112]:

```
#test
lis1T =()
counter = 0
for sent in tok_nameT:
    x = 0
    counter = counter + 1
    if counter == 60000:
        print(counter)
        counter = 0
    for w in sent:
        x = x +  model1T.wv[w].sum()
    lis1T = np.append(lis1T,x)
```

```
60000
60000
60000
60000
60000
60000
60000
```

In [113]:

```
#train
se1 = pd.Series(lis1)
```

In [114]:

```
#test
se1T = pd.Series(lis1T)
```

In [115]:

```
#train
x_train['name_num_w2v'] = se1.values
```

In [116]:

```
#test
x_test['name_num_w2v'] = se1T.values
```

In [117]:

```
x_train.head()
```

Out[117]:

	train_id	name	item_condition_id	category_name	brand_
496798	496798	Tan cardigan, size medium	27.514021	Women/Sweaters/Cardigan	15.638
1315605	1315605	Picture frame	26.399296	Home/Home Décor/Photo Albums & Frames	26.660
1104183	1104183	ibloom english bread	27.514021	Vintage & Collectibles/Antique/Collectibles	26.660
424705	424705	Terry's chocolate oranges	26.399296	Home/Kitchen & Dining/Coffee & Tea Accessories	26.660
145825	145825	Toms Wedge Heels Size 6	26.458888	Women/Shoes/Sandals	23.284

In [118]:

```
x_test.head()
```

Out[118]:

	train_id	name	item_condition_id	category_name	brand_name	ship
777341	777341	F/ship 4 Totoro Washi + 1 pen	26.530201	Handmade/Paper Goods/Stationery	26.821436	1
1463629	1463629	UCLA Men's Bundle + Shorts	26.530201	Women/Other/Other	43.858812	1
350669	350669	Listing for lol	26.530201	Beauty/Makeup/Lips	26.821436	1
310222	310222	25 pcs kawaii sticker flakes	26.530201	Kids/Toys/Arts & Crafts	26.821436	1
759257	759257	Chanel Mini Lipgloss Set	27.685444	Beauty/Makeup/Lips	81.530612	1

In [ ]:

```

```

In [119]:

```
xTrain = x_train.copy()
```

In [120]:

```
xTrain = xTrain.iloc[0:,[0,2,4,5,7,8,9,10,11,12]]
```

In [121]:

```
xTrain.head()
```

Out[121]:

	train_id	item_condition_id	brand_name	shipping	main_cat	subcat_1	su
496798	496798	27.514021	15.638889	0	28.8285	26.4586	27
1315605	1315605	26.399296	26.660021	0	24.576	21.7478	13
1104183	1104183	27.514021	26.660021	1	27.1579	23.9518	23
424705	424705	26.399296	26.660021	1	24.576	28.5768	29
145825	145825	26.458888	23.284133	0	28.8285	41.7589	30

In [ ]:

```

```

In [122]:

```
#test
xTest = x_test.copy()
xTest = xTest.iloc[0:,[0,2,4,5,7,8,9,10,11,12]]
xTest.head()
```

Out[122]:

	train_id	item_condition_id	brand_name	shipping	main_cat	subcat_1	s
777341	777341	26.530201	26.821436	1	18.048	11.2737	1
1463629	1463629	26.530201	43.858812	1	28.9335	25.285	2
350669	350669	26.530201	26.821436	1	19.807	18.8571	1
310222	310222	26.530201	26.821436	1	20.661	21.4706	1
759257	759257	27.685444	81.530612	1	19.807	18.8571	1

In [ ]:

```

```

In [123]:

```
xTrain.shape
```

Out[123]:

```
(726440, 10)
```

In [124]:

```
y_train.shape
```

Out[124]:

(726440, 1)

In [125]:

```
xTest.shape
```

Out[125]:

(444759, 10)

In [126]:

```
y_test.shape
```

Out[126]:

(444759, 1)

In [ ]:

```

```

In [127]:

```
xTrain['main_cat'] = pd.to_numeric(xTrain['main_cat'])
xTrain['subcat_1'] = pd.to_numeric(xTrain['subcat_1'])
xTrain['subcat_2'] = pd.to_numeric(xTrain['subcat_2'])
```

In [128]:

```
xTest['main_cat'] = pd.to_numeric(xTest['main_cat'])
xTest['subcat_1'] = pd.to_numeric(xTest['subcat_1'])
xTest['subcat_2'] = pd.to_numeric(xTest['subcat_2'])
```

In [129]:

```
xTrain.head()
```

Out[129]:

	train_id	item_condition_id	brand_name	shipping	main_cat	subcat_1
496798	496798	27.514021	15.638889	0	28.828531	26.458649
1315605	1315605	26.399296	26.660021	0	24.576033	21.747832
1104183	1104183	27.514021	26.660021	1	27.157887	23.951831
424705	424705	26.399296	26.660021	1	24.576033	28.576805
145825	145825	26.458888	23.284133	0	28.828531	41.758890

In [ ]:

```
```

In [ ]:

```
```

In [130]:

```
from sklearn import linear_model
```

Linear Regressin

In [131]:

```
regr = linear_model.LinearRegression(n_jobs=-1)
```

In [132]:

```
regr.fit(xTrain,y_train)
```

Out[132]:

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=-1, normali
ze=False)
```

In [133]:

```
from sklearn.metrics import mean_squared_error
```

In [134]:

```
y_pred = regr.predict(xTest)
```

In [135]:

```
print("Mean squared error: %.2f"% mean_squared_error(y_test, y_pred))
```

```
Mean squared error: 1108.79
```

In [136]:

```
from sklearn.metrics import accuracy_score
```

In [137]:

```
from sklearn.metrics import mean_squared_error, r2_score
```

In [138]:

```
print('Variance score: %.2f' % r2_score(y_test, y_pred))
```

```
Variance score: 0.27
```

In [ ]:

```
```

In [ ]:

```
```

In [160]:

```
def rmsle(y_test, y_pred):
    assert len(y_test) == len(y_pred)
    return np.sqrt(np.mean(np.power(np.log1p(y_test+1)-np.log1p(y_pred+1), 2)))
```

In [161]:

```
rmsle(y_test,y_pred)
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/ipykernel_lau
ncher.py:3: RuntimeWarning: invalid value encountered in log1p
  This is separate from the ipykernel package so we can avoid doing
imports until
```

Out[161]:

```
price    0.654986
dtype: float64
```

In [141]:

```
def symm_mean_absolute_percentage_error(y_true, y_pred):
    y_true, y_pred = np.array(y_true), np.array(y_pred)
    return np.mean(np.abs((y_true - y_pred)) / (np.abs(y_true)+np.abs(y_pred)))
* 200
```

In [142]:

```
err = symm_mean_absolute_percentage_error(y_test, y_pred)
err
```

Out[142]:

```
53.419021414930135
```

In [ ]:

```
```

Random Forests

In [143]:

```
from sklearn.ensemble import RandomForestRegressor
```

In [144]:

```
regr = RandomForestRegressor(max_depth=2, random_state=0,n_jobs=-1)
```

In [145]:

```
regr.fit(xTrain,y_train)
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/ipykernel_lau
ncher.py:1: DataConversionWarning: A column-vector y was passed when
a 1d array was expected. Please change the shape of y to (n_sample
s,), for example using ravel().
  """Entry point for launching an IPython kernel.
```

Out[145]:

```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=2,
           max_features='auto', max_leaf_nodes=None,
           min_impurity_split=1e-07, min_samples_leaf=1,
           min_samples_split=2, min_weight_fraction_leaf=0.0,
           n_estimators=10, n_jobs=-1, oob_score=False, random_state
=0,
           verbose=0, warm_start=False)
```

In [146]:

```
print(regr.feature_importances_)
```

```
[ 0.          0.          0.64048236  0.          0.          0.
  0.35951764  0.          0.          0.        ]
```

In [147]:

```
y_predictRf = regr.predict(xTest)
```

In [148]:

```
y_predictRf.shape
```

Out[148]:

```
(444759,)
```

In [162]:

```
rmsle(y_test['price'],y_predictRf)
```

Out[162]:

```
0.67816872616227242
```

In [150]:

```
errRf = symm_mean_absolute_percentage_error(y_test.price, y_predictRf)
errRf
```

Out[150]:

```
55.750520880156209
```

In [ ]:

```

```

In [151]:

```
import xgboost as xgb
from xgboost.sklearn import XGBClassifier
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/sklearn/cross
_validation.py:44: DeprecationWarning: This module was deprecated in
version 0.18 in favor of the model_selection module into which all t
he refactored classes and functions are moved. Also note that the in
terface of the new CV iterators are different from that of this modu
le. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)
```

In [152]:

```
from sklearn import cross_validation, metrics
from sklearn.grid_search import GridSearchCV
```

```
/home/ajetias129/anaconda3/lib/python3.5/site-packages/sklearn/grid_
search.py:43: DeprecationWarning: This module was deprecated in vers
ion 0.18 in favor of the model_selection module into which all the r
efactored classes and functions are moved. This module will be remov
ed in 0.20.
  DeprecationWarning)
```

In [153]:

```
xTrain.dtypes
```

Out[153]:

```
train_id              int64
item_condition_id    float64
brand_name           float64
shipping               int64
main_cat             float64
subcat_1             float64
subcat_2             float64
isBrandNull            int64
descp_num_w2v        float64
name_num_w2v         float64
dtype: object
```

In [154]:

```
xTrain.head()
```

Out[154]:

	train_id	item_condition_id	brand_name	shipping	main_cat	subcat_1	
496798	496798	27.514021	15.638889	0	28.828531	26.458649	2
1315605	1315605	26.399296	26.660021	0	24.576033	21.747832	1
1104183	1104183	27.514021	26.660021	1	27.157887	23.951831	2
424705	424705	26.399296	26.660021	1	24.576033	28.576805	2
145825	145825	26.458888	23.284133	0	28.828531	41.758890	3

In [155]:

```
xTrain = xTrain.loc[:,~xTrain.columns.duplicated()]
```

In [156]:

```
xTest = xTest.loc[:,~xTrain.columns.duplicated()]
```

In [157]:

```
#gbm = xgb.XGBClassifier(max_depth=3, n_estimators=50, learning_rate=0.05)
```

In [158]:

```
#gbm.fit(xTrain, y_train)
```

In [163]:

```
xTest.dtypes
```

Out[163]:

```
train_id              int64
item_condition_id    float64
brand_name           float64
shipping               int64
main_cat             float64
subcat_1             float64
subcat_2             float64
isBrandNull            int64
descp_num_w2v        float64
name_num_w2v         float64
dtype: object
```

In [164]:

```
data_train = xgb.DMatrix(xTrain, label=y_train)
data_valid = xgb.DMatrix(xTest, label=y_test)


watchlist = [(data_train, 'train'), (data_valid, 'test')]

xgb_params = {'min_child_weight': 20,
              'eta': 0.013,
              'colsample_bytree': 0.45,
              'max_depth': 16,
            'subsample': 0.88,
              'lambda': 2.07,

              'booster' :
              'gbtree',
              'silent': 1,
            'eval_metric': 'mae',
              'objective': 'reg:linear'}
```

In [165]:

```
model_xgb = xgb.train(xgb_params, data_train, 2000, watchlist, early_stopping_ro
unds=20, verbose_eval=50)
```

```
[0]     train-mae:25.8235       test-mae:26.0068
Multiple eval metrics have been passed: 'test-mae' will be used for
early stopping.

Will train until test-mae hasn't improved in 20 rounds.
[50]    train-mae:15.056        test-mae:15.6288
[100]   train-mae:12.6247       test-mae:13.9716
Stopping. Best iteration:
[122]   train-mae:12.3227       test-mae:13.913
```

In [166]:

```
model_xgb = xgb.train(xgb_params, data_train, 2000, watchlist, early_stopping_ro
unds=20, verbose_eval=50)
```

```
[0]     train-mae:25.8235       test-mae:26.0068
Multiple eval metrics have been passed: 'test-mae' will be used for
early stopping.

Will train until test-mae hasn't improved in 20 rounds.
[50]    train-mae:15.056        test-mae:15.6288
[100]   train-mae:12.6247       test-mae:13.9716
Stopping. Best iteration:
[122]   train-mae:12.3227       test-mae:13.913
```

In [ ]:

```

```

In [167]:

```
data_test = xgb.DMatrix(xTest)
test_predict = model_xgb.predict(data_test)
```

In [168]:

```
errXgb = symm_mean_absolute_percentage_error(y_test.price, test_predict)
errXgb
```

Out[168]:

48.635320474962761

In [169]:

```
rmsle(y_test['price'],test_predict)
```

Out[169]:

0.5944274668413263

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: