"Community Name" is All You Need? What Feature Really Matters

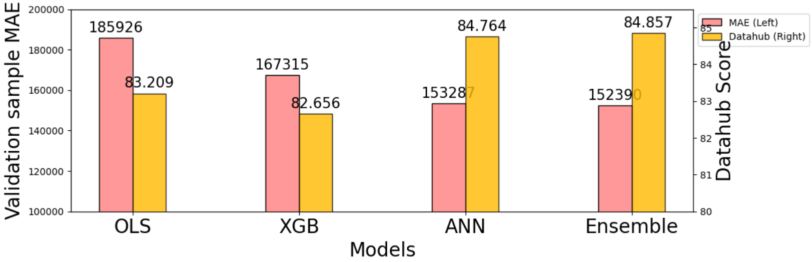
Hongming Liang
2022201480

2025/06/12

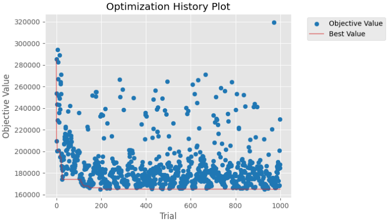
Scores

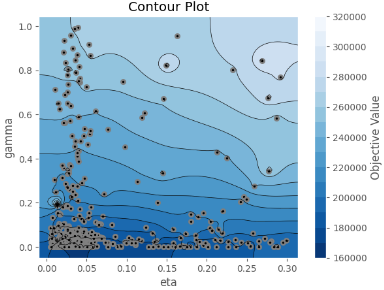
Table: RMSE for Models

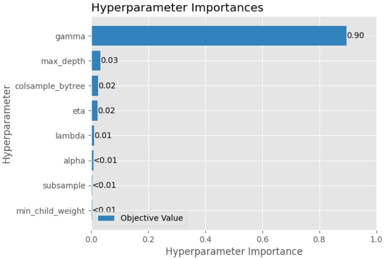
Model	In sample	Out of sample	Datahub Score
OLS	467447	542005	83.209
RF	-	-	80.538
XGB	182028	581316	82.656
Embedded ANN	-	-	82.435
ANN	298844	491192	84.764
Ensemble	301843	479520	84.857

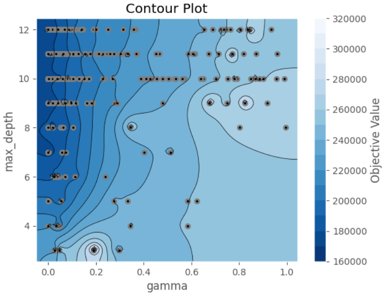


XGB with Optuna. See xgb_1.ipynb

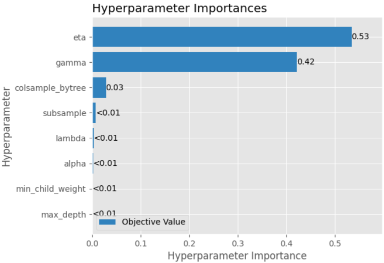


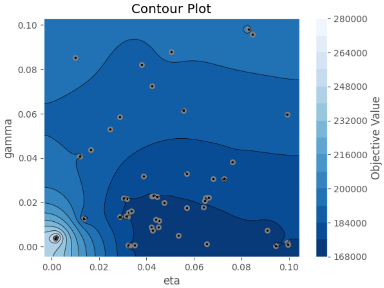




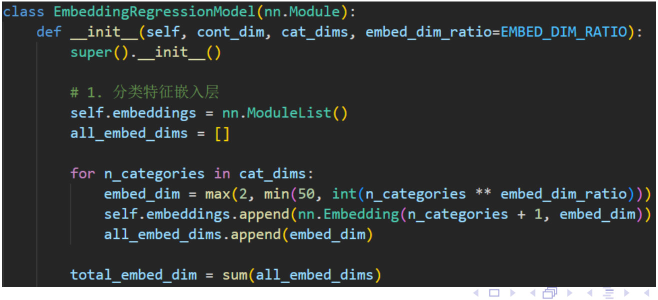


- XGB with Optuna. See xgb_2.ipynb

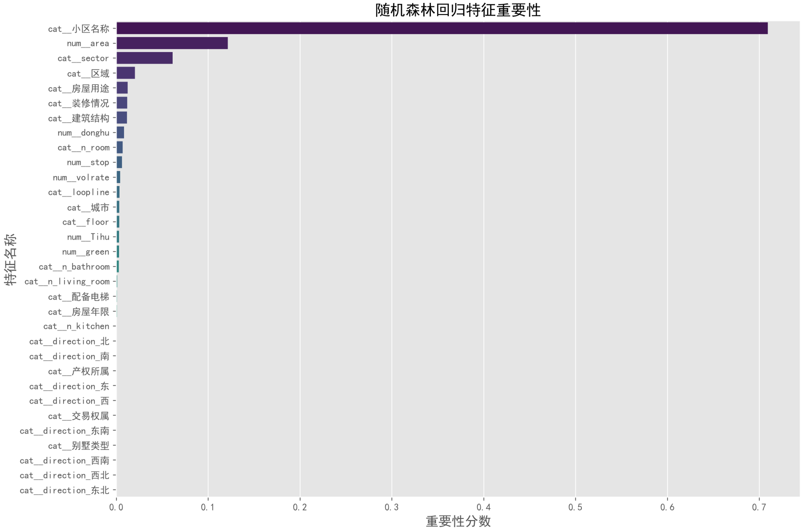




- Ann with embedded category feature. See nn_embbedded.ipynb



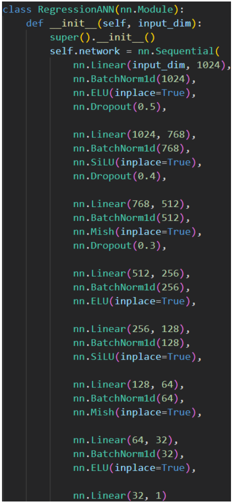
```
class EmbeddingRegressionModel(nn.Module):
    def __init__(self, cont_dim, cat_dims, embed_dim_ratio=EMBED_DIM_RATIO):
        super().__init__()

        # 1. 分类特征嵌入层
        self.embeddings = nn.ModuleList()
        all_embed_dims = []

        for n_categories in cat_dims:
            embed_dim = max(2, min(50, int(n_categories ** embed_dim_ratio)))
            self.embeddings.append(nn.Embedding(n_categories + 1, embed_dim))
            all_embed_dims.append(embed_dim)

        total_embed_dim = sum(all_embed_dims)
```



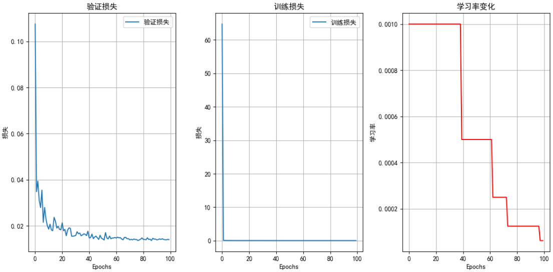
OLS	(1)	(2)	(3)	(4)	(5)
Varible	Area	(1)+City	(1)+Community	(3)+City+Region	(4)+Sector
Scores	25.892	53.814	73.091	81.235	82.333

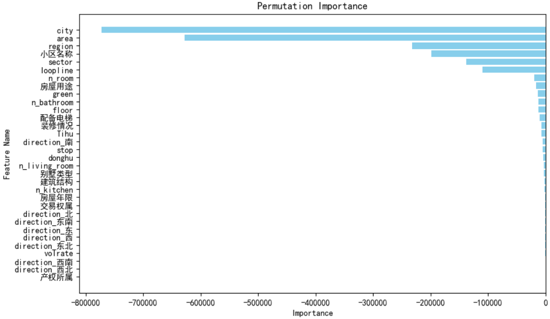
Best Model: ANN

See ann_best.ipynb

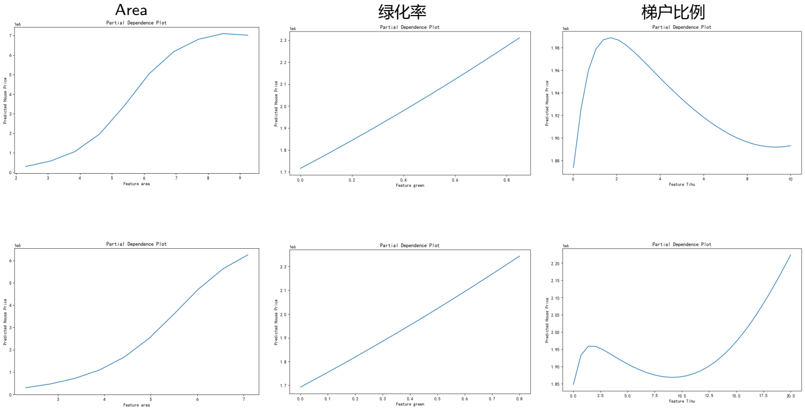


```
class RegressionANN(nn.Module):
    def __init__(self, input_dim):
        super().__init__()
        self.network = nn.Sequential(
            nn.Linear(input_dim, 1024),
            nn.BatchNorm1d(1024),
            nn.ELU(inplace=True),
            nn.Dropout(0.5),

            nn.Linear(1024, 768),
            nn.BatchNorm1d(768),
            nn.SiLU(inplace=True),
            nn.Dropout(0.4),

            nn.Linear(768, 512),
            nn.BatchNorm1d(512),
            nn.Mish(inplace=True),
            nn.Dropout(0.3),

            nn.Linear(512, 256),
            nn.BatchNorm1d(256),
            nn.ELU(inplace=True),

            nn.Linear(256, 128),
            nn.BatchNorm1d(128),
            nn.SiLU(inplace=True),

            nn.Linear(128, 64),
            nn.BatchNorm1d(64),
            nn.Mish(inplace=True),

            nn.Linear(64, 32),
            nn.BatchNorm1d(32),
            nn.ELU(inplace=True),

            nn.Linear(32, 1)
```





- PDP of ANN



- Ensemble model: OLS: XGB: ANN = 1: 0: 7.1 (Using Optuna)
- Insights: Community Name is (almost) all you need. Location information is the most important.
- Further exploration: (1)LLM? (2)ResNet? (3)coordinates(kNN)? (4)transaction time? (5)Macro features?