Real Estate Price Class Estimation

Machine Learning In Finance – Group 1

Alexander Beck | Joel Blank | Pascal Müntener | Arthur Pavlics | Kenan Öztürk

16.04.2023

Handling The Data

Features

- Feature selection
- Make composite features
- Dummies categorical features
- Scale all numericals
- Convert strings to numbers

NAN Values

- No «real» NaN values

```
 1) GrLivArea 0.1113
 2) TotalBsmtSF 0.0909
 3) OverallQual 0.0740
 4) LotArea 0.0516
 5) AgeWhenSold 0.0493
 6) BsmtFinSF1 0.0469
 7) AgeSinceRemod 0.0418
 8) 2ndFlrSF 0.0413
 9) ExterQual 0.0385
10) KitchenQual 0.0342
11) BsmtQual 0.0338
12) GarageCars 0.0319
13) OverallCond 0.0233
14) FireplaceQu 0.0205
15) HeatingQC 0.0174
16) BedroomAbvGr 0.0164
17) MSSubClass_30.0 0.0081
18) MSSubClass_60.0 0.0081
```

Data Imbalance

Problem
- Big imbalance

Solution
- Fuse class 4 and 5
- Upsampling
- Downsampling
- Tomek Links

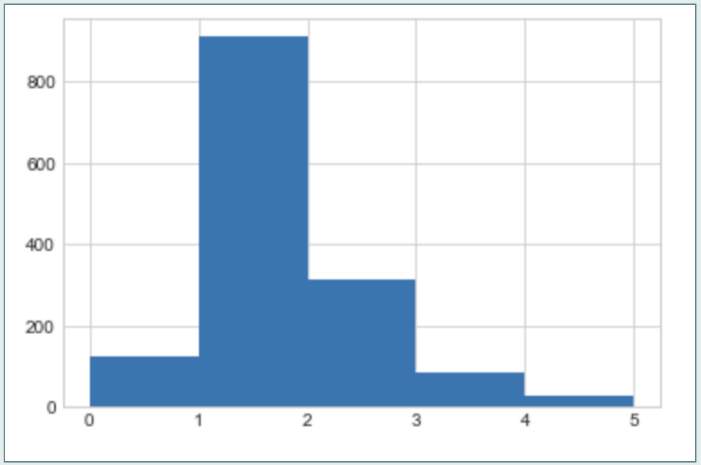
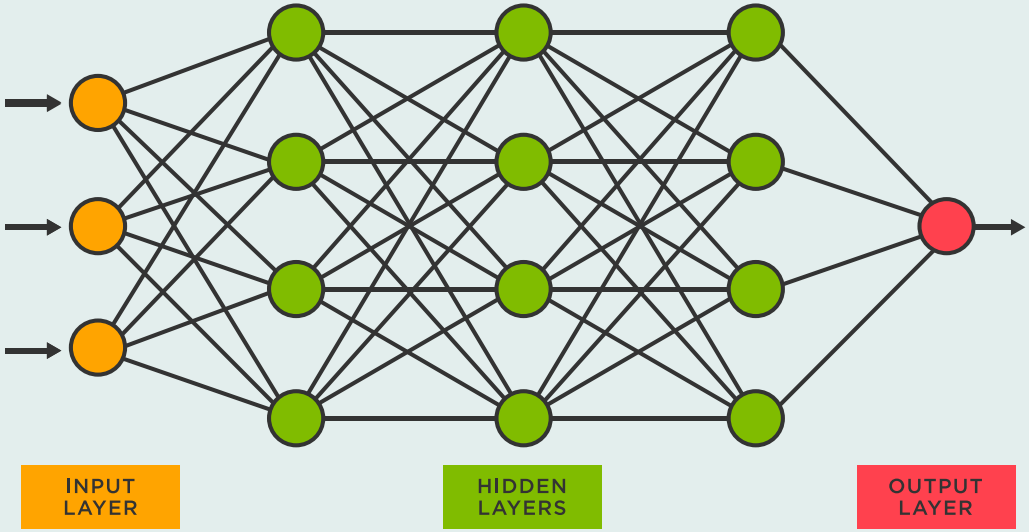


Image 2

Neural Net

Powerful tool, but not enough data

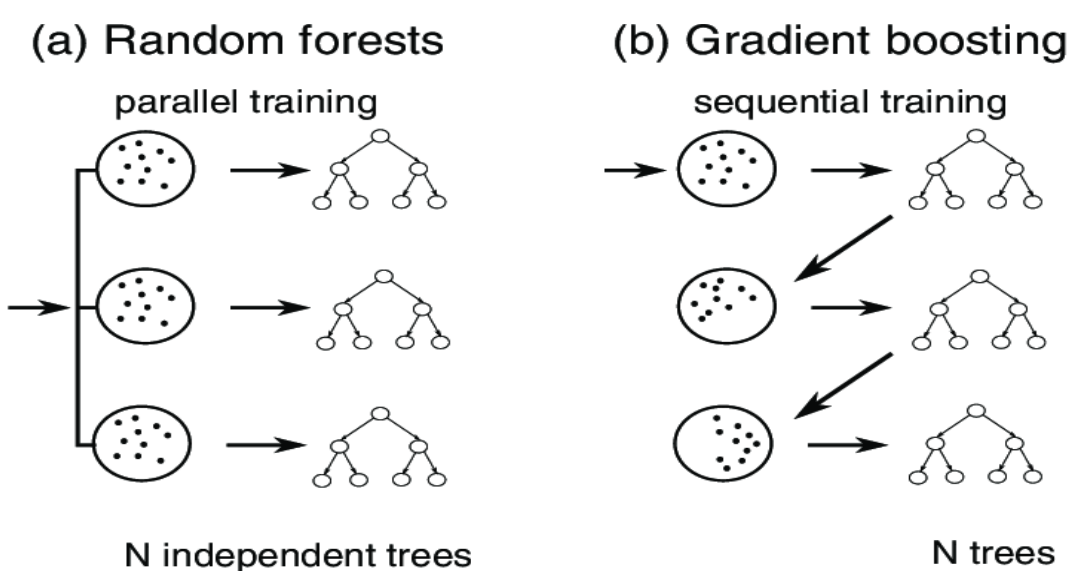
Accuracy: 78.2%



Boosted Random Forest

Best performing

Accuracy AdaBoost: 86%
Accuracy Gradient Boost: 88.7%



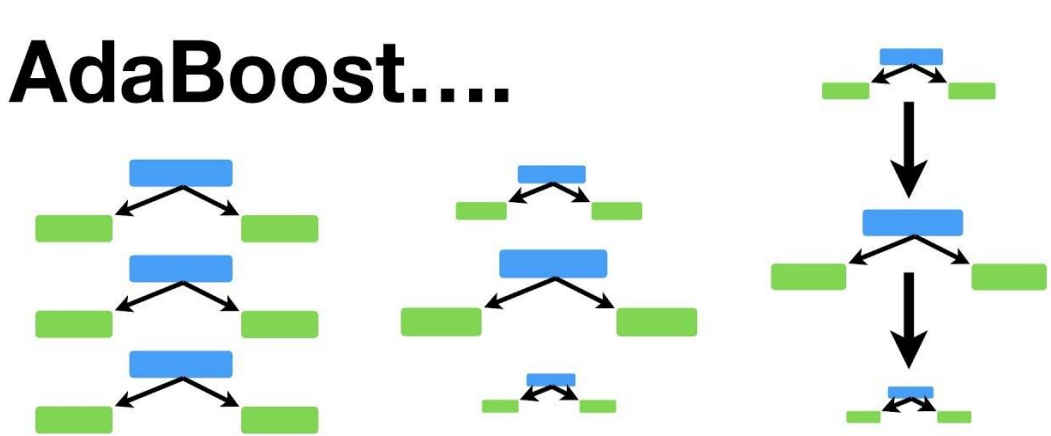


Image: 4, 5

Model Comparison

01

Gradient Boost

Accuracy: 88.7%
F1 Score: 0.88

```
GRADIENT BOOST:
 [[  8   7   0   0   0]
 [  1 184   5   0   1]
 [  0   6  54   3   0]
 [  0   0   5  11   2]
 [  0   0   0   3   3]]
```

02

Random Forest

Accuracy: 88.1%
F1 Score: 0.88

```
RANDOM FOREST:
 [[  8   7   0   0   0]
 [  0 186   4   0   1]
 [  0  10  49   4   0]
 [  0   0   6  12   0]
 [  0   0   0   3   3]]
```

03

AdaBoost

Accuracy: 86%
F1 Score: 0.85

Images 6,7,8

```
ADABOOST:
 [[  7   8   0   0   0]
 [  0 184   6   0   1]
 [  0  10  48   5   0]
 [  0   0   8  10   0]
 [  0   0   0   3   3]]
```

Model Comparison

01

Decision Trees

Accuracy: 81.6%
F1 Score: 0.82

```
DECISION TREES:
 [[  9   6   0   0   0]
 [  9 170  11   0   1]
 [  0  13  48   2   0]
 [  0   0   8   9   1]
 [  0   0   0   3   3]]
```

02

Neural Net

Accuracy: 78.2%
F1 Score: 0.77

```
NEURAL NET:
 [[  7   8   0   0   0]
 [  1 181   8   1   0]
 [  0  23  34   6   0]
 [  0   1  10   7   0]
 [  0   0   0   6   0]]
```

03

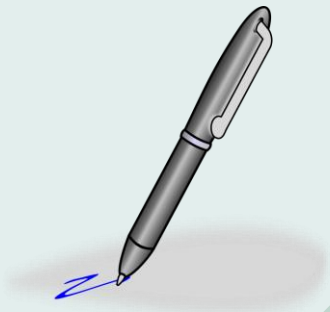
Minimal Working Example

Accuracy Rand. Forest
82%

Images 9, 10

Conclusion

- Baseline to beat: 62.33% (always guess most common class)
- All models beat this comfortably
- NNs may not be the best tool for this specific task
  -> Hyperparameteroptimization



BIBLIOGRAPHICAL REFERENCES

References

Beck, A., Blank, J., Müntener, P., Pavlics, A. & Öztürk, K. (2023) Jupyter Notebooks – ML GROUP PROJECT.

Schmitt, T. (2019). house_prices [Datensatz]. OpenML. https://www.openml.org/search?type=data&sort=runs&id=42165&status=active

Zimmermann, B. (2023). Jupyter Notebooks – Group Project: Minimal Working Expamle.

Images

Image 1,6,7,8,9,10: Beck, A., Blank, J., Müntener, P., Pavlics, A. & Öztürk, K. (2023) Jupyter Notebooks – ML GROUP PROJECT.

Image 2: Zimmermann, B. (2023). Jupyter Notebooks – Group Project: Minimal Working Expamle.

Image 3: TIBCO Software. (o. J.). *What is a Neural Network?*. Abgerufen am 14. April 2023 von https://www.tibco.com/reference-center/what-is-a-neural-network

Image 4: ResearchGate. (o.J.). *FIG. 1. Comparison between (a) random forest and (b) gradient boosting....* Abgerufen am 14. April 2023 von

Image 5: StatQuest with Josh Starmer. (2019). *AdaBoost, Clearly Explained*. Abgerufen am 14. April 2023 von https://www.youtube.com/watch?v=LsK-xG1cLYA

Image 11: Clipartix. (o. J.). *Ballpoint pen clipart free clipart images.* Abgerufen am 14. April 2023 von https://clipartix.com/pen-clipart-image-22207/

Slides: Slidesgo. (o. J.). *The Evolution of Invention in Canada Thesis.* Abgerufen am 14. April 2023 von https://slidesgo.com/theme/the-evolution-of-invention-in-canada-thesis#search-tech&position-34&results-354&rs=search