Sidarth Wadhwa
Rui Maki

Homework 1

CS5785
September 17, 2021

Programming Exercises

Part I. The Housing Prices

1. Joined competition and downloaded datasets.
2. Three continuous features in the dataset: Neighborhood, TotalBsmtSF, GrLivArea. Three categorical variables in the dataset: Street, PavedDrive, LandSlope

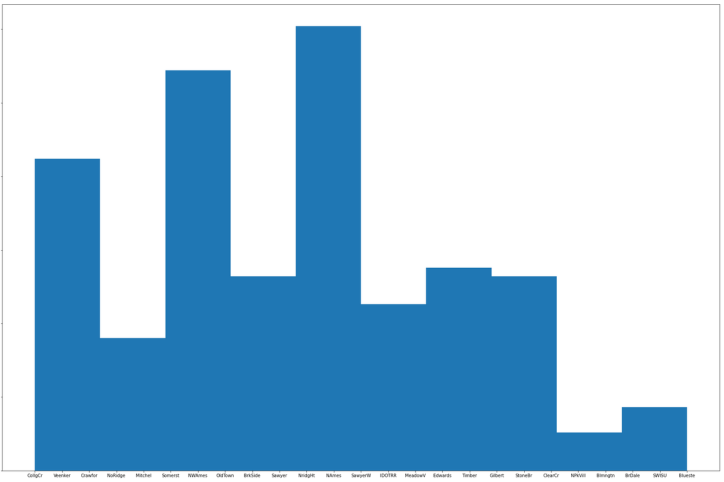


Figure 1: Histogram for Neighborhood (Categorical)

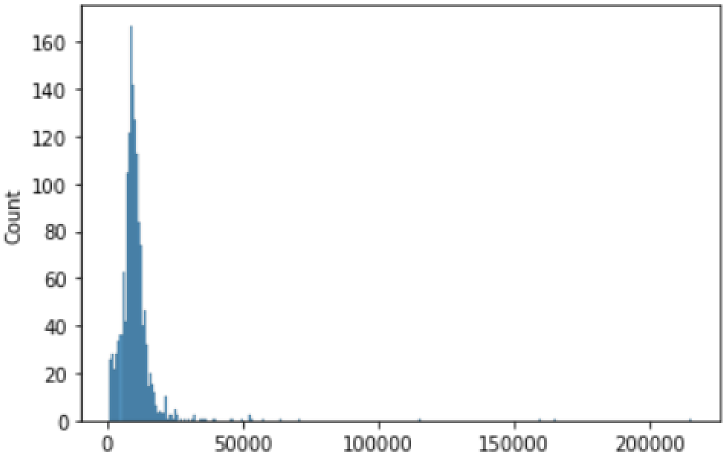
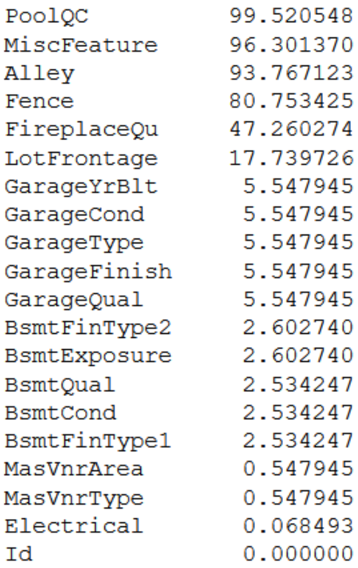


Figure 2: Histogram for LotArea (Numerical)

3. For pre-processing, we first find the percentages of missing values in the train dataset. Threshold for missing values assumed to be 10%. If percentage is over the threshold, dropping the column is considered. Since the top 6 in the above table (all above 10%) are not features that would have a high correlation with the price of the house, we can safely drop the same.



```
PoolQC          99.520548
MiscFeature     96.301370
Alley           93.767123
Fence           80.753425
FireplaceQu     47.260274
LotFrontage     17.739726
GarageYrBlt      5.547945
GarageCond       5.547945
GarageType       5.547945
GarageFinish     5.547945
GarageQual       5.547945
BsmtFinType2     2.602740
BsmtExposure     2.602740
BsmtQual         2.534247
BsmtCond         2.534247
BsmtFinType1     2.534247
MasVnrArea       0.547945
MasVnrType       0.547945
Electrical       0.068493
Id               0.000000
```

Figure 3: Percentages of missing values

Following from the above, the features that are dropped are: 'PoolQC', 'MiscFeature', 'Alley', 'Fence', 'FireplaceQu', 'LotFrontage'. Furthermore, in the interest of time, the attributes are not the primary concerns that a person looking to purchase considers. Furthermore, there are other variables that cover similar aspects that are likely to sway buyer decisions more (the total sq. ft. that the basement covers, the no. of cars that can be stored in the garage). Hence, these features are dropped as well. Masonry Veneer is another feature that does not seem like it would have an impact, which means that it too is going to be dropped. Additionally, 'Id' is column that is rather redundant for training purposes and is hence dropped.
The last remaining feature with a non-zero percentage is 'Electrical'. Since this is a more important feature and the missing values account for such a small percentage, we can just drop the missing values and keep the feature.
Next, we move onto segregating and normalizing the numerical and categorical parts of the dataset. The numerical subset of features is normalized, while the categorical variables are encoded using either One-Hot Encoding or Label Encoding.

4. Features that can use One-Hot should have only a limited number of possible values that they take, in order to ensure that the dimesionality is not increased by too much.A such, some of the possible candidate features for OHE are: 'Street', 'LandS-

lope', 'PavedDrive'. we choose to go ahead and encode 'Street'.

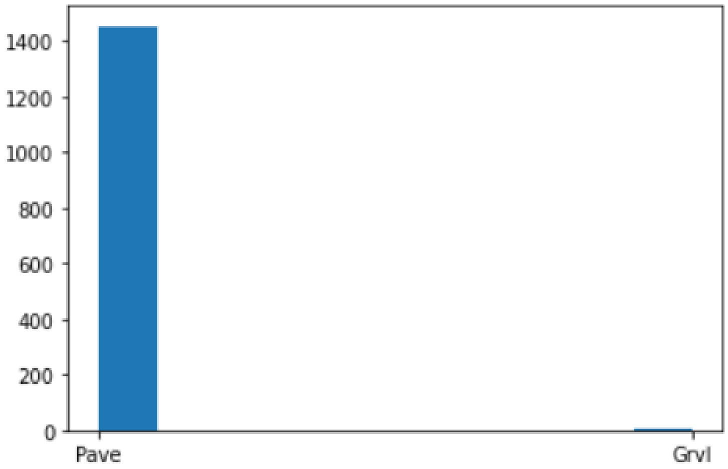


Figure 4: Street before OHE

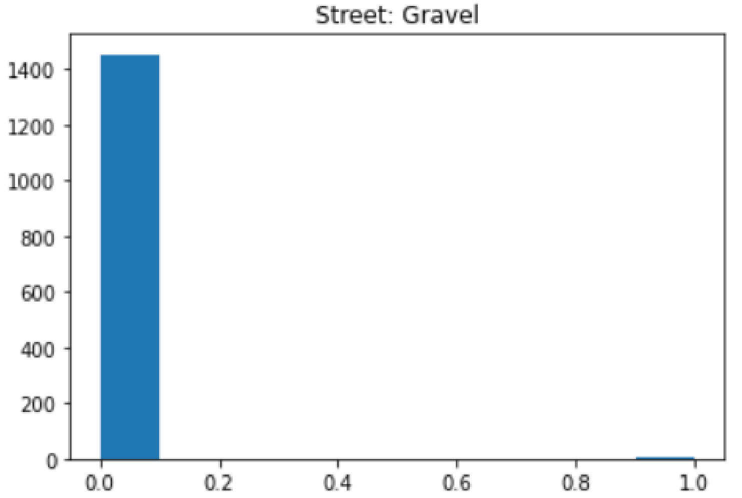


Figure 5: Street after OHE

Ordinal values can be encoded with integers that represent each of the categories. We can encode the following ordinal variables using LabelEncoder. 'ExterQual', 'ExterCond', 'HeatingQC', 'CentralAir', 'Electrical', 'KitchenQual', 'Functional', 'LotShape', 'Utilities', 'BldgType', 'HouseStyle', 'LotConfig'.

5. Implemented in the ipynb file.

6. Trained and output generated in the ipynb file. The resultant csv is the submitted to Kaggle to get a score of 0.66603 (Fig.10- at the end).

Part II. The Titanic Disaster

1. Joined competition and downloaded datasets.

2. Among all the features in the dataset, 'Sex' seeems to be a major feature in the likelihood of a passenger's survival, owing to the fact that the survival rate of females is much higher than that of males. As shown in Fig. 6, there is a curious relationship that is apparent between the general population of the ship and the survival figures when inspected in the context of sex. This is propbably of significance to our model, and we hence include 'Sex' in the columns we use to train the model.

   Next, we see that the survival rates across the 'Pclass' attribute vary quite a bit across the different classes (Fig. 7), with them going down as we go down the hierarchy of classes on the ship. This, again, is an indicator that the class a passenger was in seems to have an impact on how likely they were to survive. Therefore, we include this in the model.

   'Age' (Fig. 8) is another attribute that ought to determine whether the people survived or not. This is confirmed by graphing the attribute. However, this attribute has 177 missing values that we need to fill in. Filling it with general averages, medians, etc. would be a possible way to go about doing the same, but it would not yield the most accurate values. Hence, we decide to go for a more 'tailored' way to approach filling in the missing values, i.e, we decide to fill in the missing values with the median of the ages of the population grouped by 'Pclass' and 'Sex'. This is expected to lead to a more realistic ballpark figure than the above-mentioned approaches.

   Finally, I also choose to include 'Parch' and 'SibSp' attributes in my model's training. I do this because the former indicates the number of family members a given individual had on the train, while the latter indicates the number of siblings. Both of these ought to be considered when trying to understand the survivability of the disaster for a given person, since peers do matter in such a crisis.

3. The model is trained and the predictions are made for the test set. These are then submitted to Kaggle for a score of 0.77272 (Fig.11- at the end).

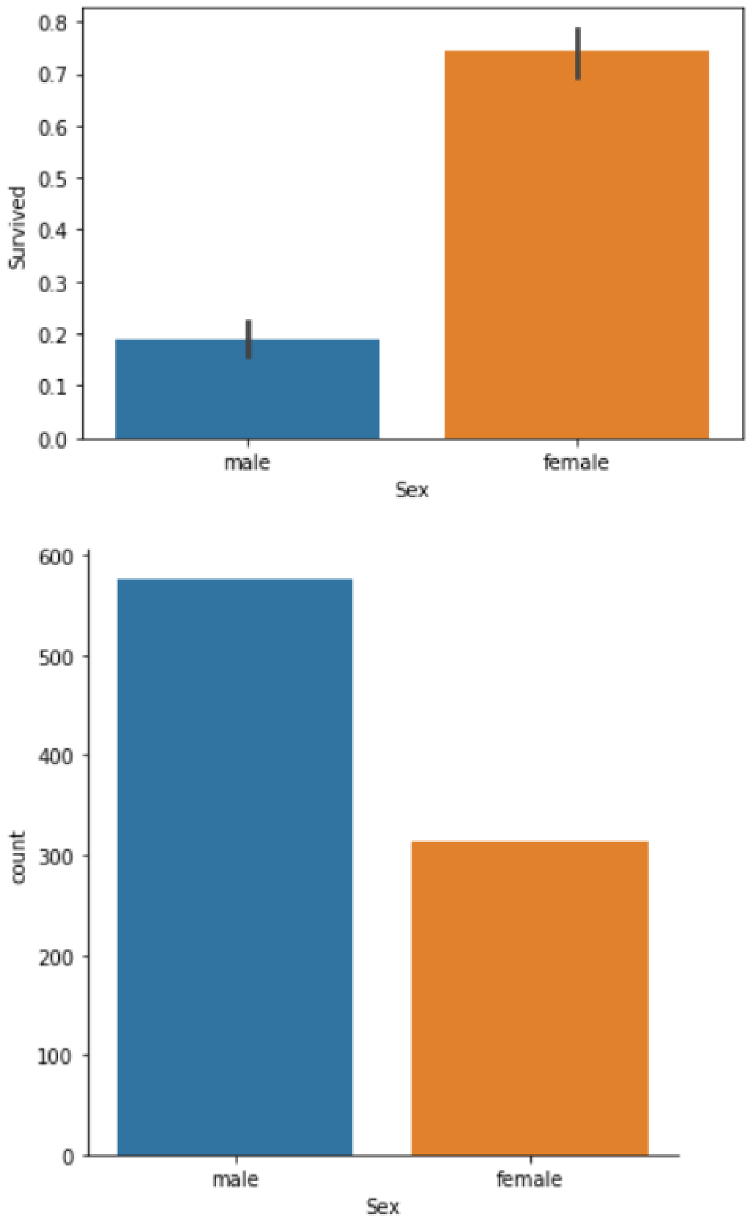


Figure 6: Sex attribute in the dataset

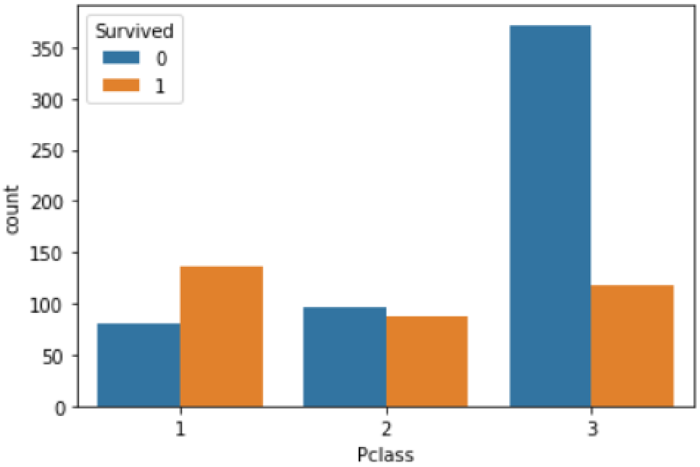


Figure 7: Survival numbers grouped by the Pclass attribute

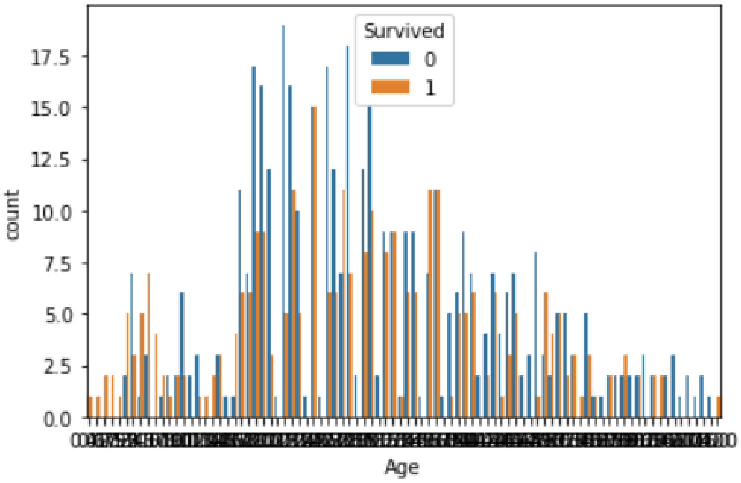


Figure 8: Sex attribute in the dataset

Written Exercises

1. To Prove:

$$\arg\max_{\theta}[E_{p(x,y)}[\log(p_{\theta}(y|x))]] = \arg\min_{\theta}[E_{p(x)}[KL(p(y|x)||p_{\theta}(y|x))]] \tag{1}$$

KL-Divergence:

$$KL(p(x)||q(x)) = E_{x\sim p(x)}[\log(p(x)) - \log(q(x))] \tag{2}$$

Proof:

$$\begin{aligned}
\arg\max_{\theta}[E_{p(x,y)}[\log(p_{\theta}(y|x))]] &= \arg\min_{\theta}[E_{p(x)}[E_{p(y|x)}[\log(p(y|x)) - \log(p_{\theta}(y|x))]]] && (3)\\
&= \arg\min_{\theta}[E_{p(x)}[E_{p(y|x)}[\log(p(y|x))]] - E_{p(x)}[E_{p(y|x)}[\log(p_{\theta}(y|x))]]]] && (4)\\
&= \arg\min_{\theta}[[E_{p(x,y)}[\log(p(y|x))]] - [E_{p(x,y)}[\log(p_{\theta}(y|x))]]] && (5)\\
&= \arg\max_{\theta}[E_{p(x,y)}[\log(p_{\theta}(y|x))]]. && (6)
\end{aligned}$$

Here, in equation (3), we utilise the definition of KL Divergence. In (4), we use the linearity of expectations. Then, using Bayes' Theorem, we're able to combine the two expectation operators in (5). At this stage, we can see that it is only the second term of the equation that depends on θ. Hence, we can drop the other term and arrive at our conclusion

2. (a) $\sigma(a) = \dfrac{1}{1+e^{-a}}$

$$\begin{aligned}\frac{d}{da}\sigma(x) &= \frac{d}{da}\left[\frac{1}{1+e^{-a}}\right] \\ &= \frac{d}{da}\left(1+\mathrm{e}^{-a}\right)^{-1} \\ &= -(1+e^{-a})^{-2}(-e^{-a}) \\ &= \frac{e^{-a}}{(1+e^{-a})^2} \\ &= \frac{1}{1+e^{-a}} \cdot \frac{e^{-a}}{1+e^{-a}} \\ &= \frac{1}{1+e^{-a}} \cdot \frac{(1+e^{-a})-1}{1+e^{-a}} \\ &= \frac{1}{1+e^{-a}} \cdot \left(\frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}\right) \\ &= \frac{1}{1+e^{-a}} \cdot \left(1 - \frac{1}{1+e^{-a}}\right) \\ &= \sigma(a)\cdot(1-\sigma(a))\end{aligned}$$

(b)

$$\begin{aligned}\frac{d}{d\theta}\left[l(\theta)\right] &= \frac{d}{d\theta}\left[y\log\sigma(\theta^T x)\right] + \frac{d}{d\theta}\left[(1-y)\log(1-\sigma\theta^T x)\right] \\ &= \frac{y}{\sigma(\theta^T x)}\frac{d}{d\theta}\sigma(\theta^T x) + \frac{1-y}{1-\sigma(\theta^T x)}\frac{d}{d\theta}(-\sigma(\theta^T x)) \\ &= \frac{d}{d\theta}\sigma(\theta^T x)[\frac{y}{\sigma(\theta^T x)} - \frac{1-y}{1-\sigma(\theta^T x)}] \\ &= \sigma(\theta^T x)[1-\sigma(\theta^T x)]x[\frac{y}{\sigma(\theta^T x)} - \frac{1-y}{1-\sigma(\theta^T x)}] \\ &= (y-\sigma(\theta^T x))x\end{aligned}$$

3. Points: (-1,-1), (-1,0), (0,-1), (0,0), (0,1), (1,0), (1,1)

(a) The plot of the points is shown in Fig.9. As for the best fit line, my assumption would be that it would be a line of Slope +1, going through the Origin (Intercept). This is assuming that the points outside the line will be at minimum cumulative distance from it in that case.

(b) For this question, we're going to use the following formulae to obtain the slope and intercept of the lines:

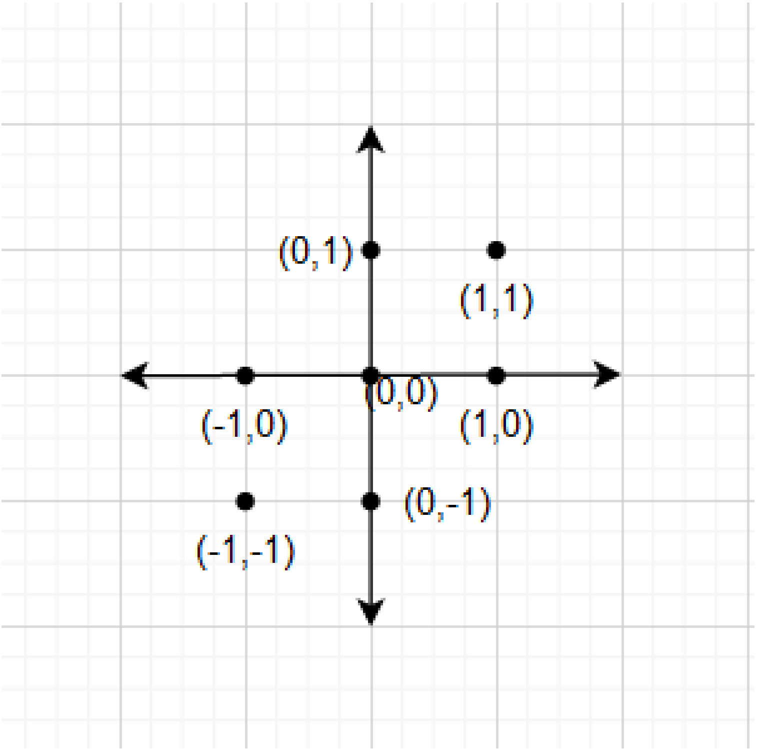


Figure 9: Plot of the Points

$$MSE = \frac{1}{N}\sum_{i=0}^{N}(y_i - y)^2$$

$$m = \frac{\overline{xy} - \overline{x}\overline{y}}{\overline{x^2} - (\overline{x})^2}, b = \overline{y} - m\overline{x}$$

Now to obtain values of the individual terms in the formula by substituting the values of the points.

$$\overline{x} = \frac{(-1) + (-1) + 0 + 0 + 0 + 1 + 1}{7} = 0$$

$$\overline{x^2} = \frac{(-1)^2 + (-1)^2 + 0 + 0 + 0 + 1^2 + 1^2}{7} = \frac{4}{7}$$

$$\overline{y} = \frac{(-1) + 0 + (-1) + 0 + 1 + 0 + 1}{7} = 0$$

$$\overline{xy} = \frac{(-1)(-1) + (-1)(0) + (0)(-1) + (0)(0) + (0)(1) + (1)(0) + (1)(1)}{7} = \frac{2}{7}$$

Now, for slope.

$$m = \frac{\frac{2}{7} - 0}{\frac{4}{7} - 0} = \frac{1}{2}$$

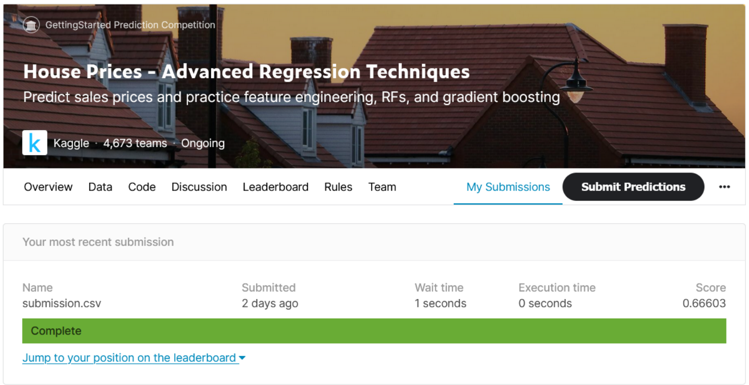
Therefore, our line is:

$$y = mx + b$$
$$y = \frac{1}{2}x$$

(c)

$$MAE = \frac{1}{N}\sum_{i=0}^{N}(|y_i - y|)$$

Now, we can see that the the MAE differs from the MSE in that there is a lack of squared terms. Instead we take the absolute value of each term in the distribution and sum it up. Keeping this quirk in mind, we can deduce that, since the points lie within the two opposite quadrants, the sum of their errors will remain the same so long as the slope of the line stays within the range of 0 to 1.
Therefore, the slope of the line $m = [0, 1]$ and $b = 0$


GettingStarted Prediction Competition
House Prices - Advanced Regression Techniques
Predict sales prices and practice feature engineering, RFs, and gradient boosting
Kaggle · 4,673 teams · Ongoing
Overview
Data
Code
Discussion
Leaderboard
Rules
Team
My Submissions
Submit Predictions
...

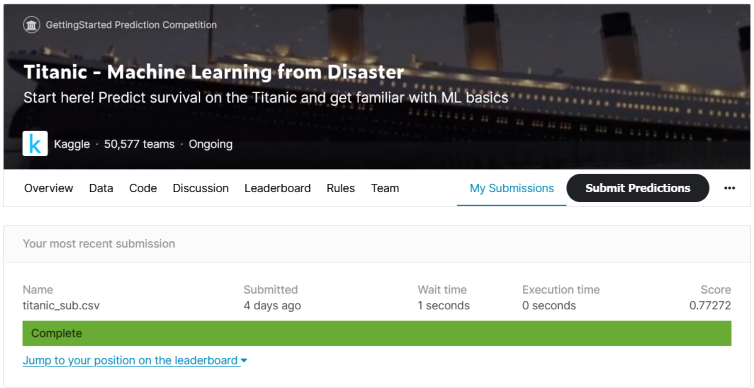
Your most recent submission

Name	Submitted	Wait time	Execution time	Score
submission.csv	2 days ago	1 seconds	0 seconds	0.66603

Complete

Jump to your position on the leaderboard

Figure 10: Kaggle submission for Housing


GettingStarted Prediction Competition
Kaggle · 50,577 teams · Ongoing
Overview
Data
Code
Discussion
Leaderboard
Rules
Team
My Submissions
Submit Predictions
...
Jump to your position on the leaderboard

Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
titanic_sub.csv	4 days ago	1 seconds	0 seconds	0.77272

Complete

Figure 11: Kaggle submission for Titanic