Predicting Lancaster City Housing Prices and Rent Status

Davis Cook, Matt Turetsky, Lynn Zhao

May 4, 2021

Contents

List of Tables

List of Figures

1 Introduction

The housing market is a fickle thing. Especially since the housing bubble, subsequent market crash of 2008, and constant increasing levels of gentrification, buyers and sellers are keen to predict the price of real estate in their neighborhood. Lancaster County, in particular Lancaster City, has seen a recent boom in state-wide and national notoriety. The county has long been famous for its Plain community that reside in the rural towns of the county. But recent interest has been placed on the City of Lancaster as a growing and culturally important urban center of Pennsylvania. The city has been called by the BBC "America's Refugee Capital" in 2017 (Strasser 2017); The New York Times, in 2019, spotlighted the unique food scene of the city, calling it "a hive of culinary diversity" (Krishna 2019).

This recent surge in the county's recognition begs the question of how this fame will affect the real estate market. This project aims to build a model that predicts the price of property, both commercial and residential, in Lancaster City using data provided by the county. Data was taken from the Lancaster County Property Tax Inquiry website (https://lancasterpa.devnetwedge.com/). The database provides property data of land sales in Lancaster County going back to 1900, although data before 2005 has been seen to be unreliable. Thus, data is pulled from All residential land sales from 2005 until 2020. After removing outlier properties that sold for less than $10,000—which typically occurs when a house is foreclosed—and greater than $1,000,000, our model uses property information from 17,337 locations in Lancaster City.

1.1 Predictors

The database offers data related to the qualities of the house, such as the price and square footage, and to the features of the owner, such as owner address. Our model uses only the intrinsic features of the property to attempt to predict the sale price of a house and whether the house is rented or owned. The predictors used in our models are displayed in Table 1.

Table 1: Summary of predictors used in model

Feature	Description
LIVING_SQFT	Total living square footage
PROPERTY_SQFT	Total property square footage
Num_structures	Number of structures per property
Num_STORIES	Number of stories in building
AGE	Year house was built
FIPS_TRACT[1]	Classifier of Census Tract Identification

Feature	Description
BASEMENT_AREA	Total area of Basement
Full.Baths	Number of full bathrooms
Number.of.Rooms	Number of rooms
Number.of.Bedrooms	Number of bedrooms
Number.of.Families	Number of families
sale_year	Year of sale between 2005 and 2020
Extra_Fixtures	Number of added fixtures
OUTDOOR_AREA	Total outdoor area of property

We also engineered several features for our model. Of note is the "logZscore" of categorical variables roof type, wall material and heating type. Given the many-leveled nature of these variables, we transformed their levels into standard deviations from the mean price of each roof/wall/heating type, averaged over each census tract and each year. Thus, each logZscore variable represents a "score" of how desirable a particular category is. We understand this method may be left with criticism, however, our data is more dirty and raw than most. We see this as an effective way of measuring an otherwise dense variable, and a fun thing to try in a non-published paper.

1.2 Research Questions and Response Variables

We proposed one regression and one classification question to answer using the predictors listed above.

1. What is the predicted sale price of a property in Lancaster City?
2. What is the predicted probability, using intrinsic features of the property, that a house is owned or rented?

We add this stipulation to our second question regarding rent status because Davis' summer research has shown that the Lancaster County Property Tax Inquiry website provides the owner's mailing address, so simply comparing the mailing address to the property's location gives a good indication if the property is owned or rented (i.e. if the two addresses are not the same, it is likely that the owner does not live in the house; thus, the property is likely rented). The responses variables are summarized in Table 2

[1]FIPS stands for Federal Identification Processing Standards. See more on the numerical identification system here.

Table 2: Summary of response variables

Response	Description
price	Sale price of property
rented	Binary class of rent status: owned = 0, rented = 1

The distribution of sale price can be seen in Figure 1.

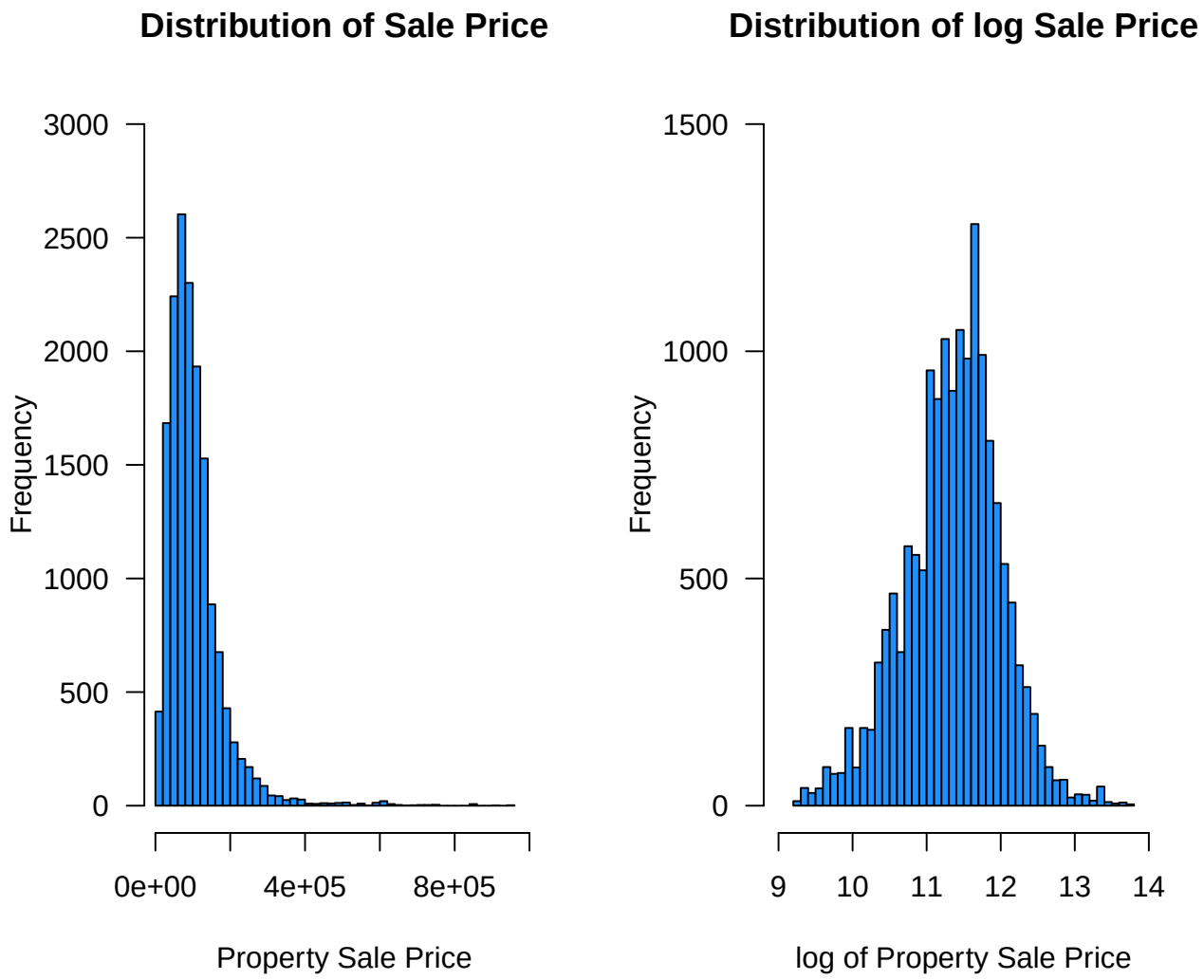


Figure 1: Histograms of sale price for Lancaster City homes

The distribution in Figure 1 shows a strongly skewed right distribution, that appears to be a log-normal distribution. Previous studies modeling property values have used the logarithm of the price instead, since the distribution is more normal. The distribution of $log(price)$ does appear to be approximately normal, as Figure 1 shows.

To get a rough geographical understanding of the rent status of homes—which intuitively could be a useful predictor— Table 3 shows the binary distribution of owned and rented houses in each census tract.

Table 3: Summary of rent status in Lancaster City

	Owned	Rented
FIPS ID: 100	23	187
FIPS ID: 200	181	816
FIPS ID: 300	367	1421
FIPS ID: 400	157	795
FIPS ID: 500	205	954
FIPS ID: 600	188	875
FIPS ID: 700	132	655
FIPS ID: 800	70	449
FIPS ID: 900	144	612
FIPS ID: 1000	287	1057
FIPS ID: 1100	651	723
FIPS ID: 1200	637	933
FIPS ID: 1400	428	1161
FIPS ID: 11701	3	13
FIPS ID: 11805	9	20
FIPS ID: 13202	83	46
FIPS ID: 13203	0	1
FIPS ID: 13301	68	154
FIPS ID: 13400	20	85
FIPS ID: 13501	5	2
FIPS ID: 13502	3	1
FIPS ID: 13503	106	76
FIPS ID: 14700	240	762
FIPS ID: NaN	6	59

2 Statistical Models

Modeling continuous and discrete response variables require different techniques. Continuous variables, such as *price*, are best predicted using regression techniques like linear regression,

regression trees, Principal Components Regression (PCR), and LASSO regression.

Discrete variables, typically coded as a binary response, are best modeled using various classification techniques. The most common method is using logistic regression, but other models include classification trees, Linear Discriminant Analysis, and K Nearest Neighbors.

The modeling techniques we used are summarized in the following two sections. For regression, we use Best Subset Selection and the Lasso. And for classification we use a Logistic Model and K-Nearest Neighbors

2.1 Regression Techniques

Our initial data exploration, and intuition, tells us there is be significant multicolinearity between our variables. We expect variables such as sqft and number of bedrooms, or sqft and basement size, to be correlated. We have many such examples of these variables, so our analysis must account for them.

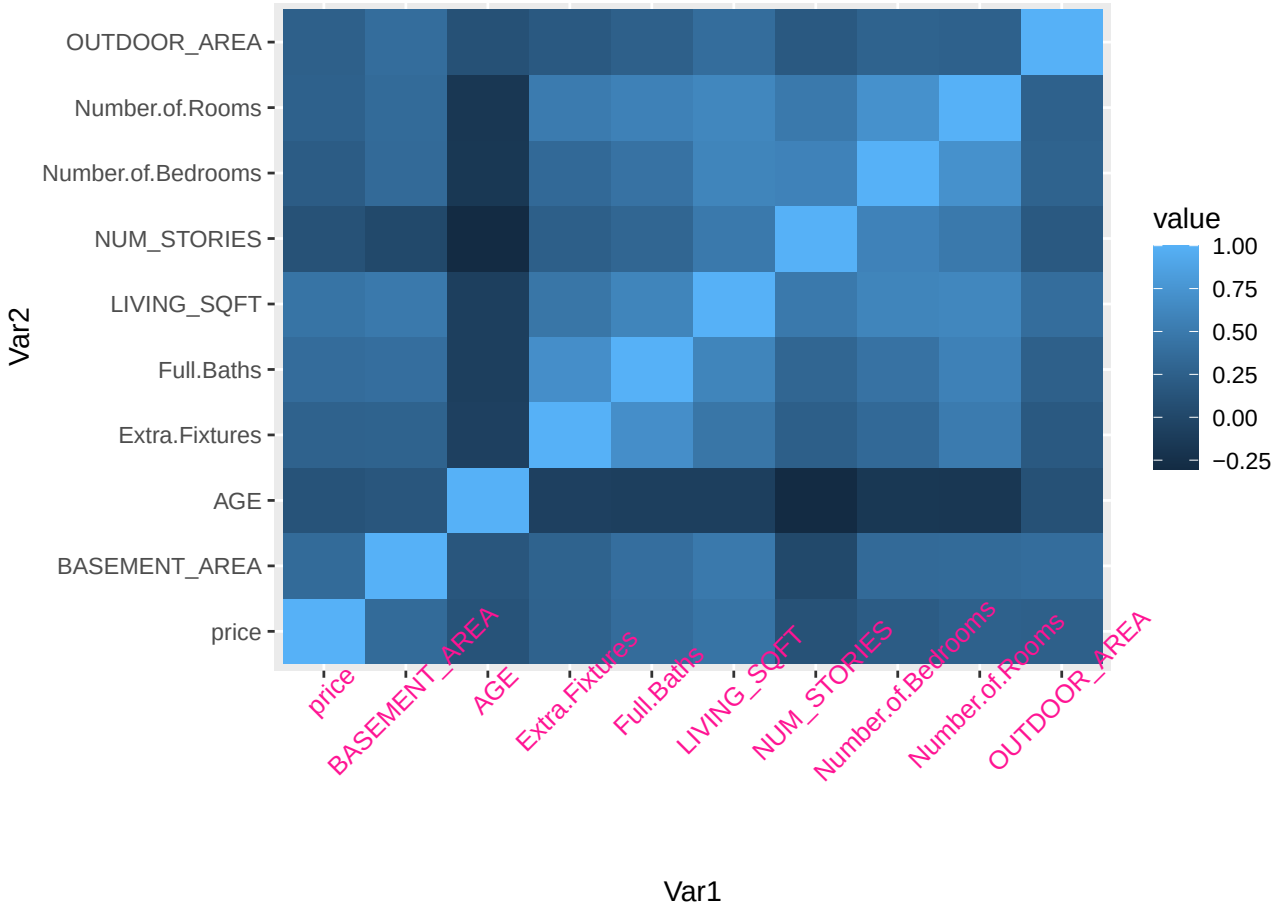


Figure 2: Correlation Heatmap

In Figure 2, we see that many of our variables have significant correlations with each other, but not particular one has a noticeably strong correlation with our response variable. Ideally, we would like to use Principal Components Regression to tease out these sources of variation, but the presence of important categorical variables in our model means that PCR is not a valid model. Thus, we continue with one subset selection model and one shrinkage model.

Forward Selection. This technique produces a multiple linear regression model of the form

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots \hat{\beta}_p X_p + \epsilon$$

where Y is the response variables, $X_1 \ldots X_p$ are the predictors, $\hat{\beta}_0 \ldots \hat{\beta}_p$ are the corresponding estimated coefficients for each predictor X_i, and ϵ is the mean zero random error term. The algorithm for Forward Selection is as follows:

1. Let M_0 be the null model which contains no predictors.
2. For $k = 1, 2, \ldots, p-1$, where p is the total number of predictors, consider all $p-k$ models that augment the predictors in M_k with one more predictor. Pick the best model and call it M_p.
3. Choose the best model among the $p-k$ models via RSS or R^2 and call it M_{k+1}
4. Select the best model among $M_0 \ldots M_p$ via cross validation error, adjusted-R^2, or some other error estimation to compare models.

We are testing the null hypothesis that

$$\beta_1 = \beta_2 = ...\beta_p = 0$$

against the alternative hypothesis that some

$$\beta_i \neq 0.$$

The relevant test statistics and p-values are available in Table 4.

Table 4: Coefficient-Level Estimates for a Model Fitted to Estimate Housing Prices.

Predictor	B	SE	t	p
(Intercept)	-42.37	1.901	-22.29	0.000
FIPS_TRACT200	-0.01	0.046	-0.29	0.769
FIPS_TRACT300	-0.06	0.044	-1.29	0.199

Predictor	B	SE	t	p
FIPS_TRACT400	-0.16	0.045	-3.56	0.000
FIPS_TRACT500	-0.05	0.043	-1.13	0.259
FIPS_TRACT600	-0.16	0.043	-3.63	0.000
FIPS_TRACT700	-0.10	0.045	-2.19	0.028
FIPS_TRACT800	-0.24	0.052	-4.69	0.000
FIPS_TRACT900	-0.23	0.049	-4.70	0.000
FIPS_TRACT1000	-0.19	0.049	-3.90	0.000
FIPS_TRACT1100	-0.04	0.046	-0.80	0.424
FIPS_TRACT1200	0.01	0.046	0.16	0.870
FIPS_TRACT1400	-0.18	0.049	-3.64	0.000
FIPS_TRACT11701	-0.29	0.216	-1.34	0.180
FIPS_TRACT11805	-0.06	0.103	-0.63	0.532
FIPS_TRACT13202	0.02	0.066	0.29	0.774
FIPS_TRACT13301	-0.15	0.058	-2.52	0.012
FIPS_TRACT13400	-0.22	0.071	-3.13	0.002
FIPS_TRACT13501	-0.02	0.184	-0.12	0.903
FIPS_TRACT13502	-0.03	0.276	-0.09	0.925
FIPS_TRACT13503	0.04	0.061	0.72	0.474
FIPS_TRACT14700	-0.27	0.052	-5.19	0.000
sale_year	0.02	0.001	25.64	0.000
AGE	0.00	0.000	15.83	0.000
Full.Baths	0.09	0.009	10.30	0.000
LIVING_SQFT	0.00	0.000	-0.05	0.960
LIVING_SQFT_sqrt	0.04	0.004	9.88	0.000
logZscoreROOF	0.15	0.010	14.56	0.000
logZscoreWALL	0.15	0.012	12.81	0.000
ABOVE_GROUND_AREA	0.00	0.000	-9.46	0.000

A full visualization of error metrics and number of variables can be found in the appendix.

While adjusted-R^2continues to rise as the number of predictors increases and Cp and BIC decrease, the change is marginal, around 10 for adjusted R^2 and Cp. So for interpretability, the highest complexity model is not chosen.

We reject the null hypothesis and find multiple statistically significant predictors of housing price. Notably, some but not all census tracts have a significant effect on logPrice. Besides sale year, the rest of the significant variables are related to the physical characteristics of the house.

Our error metrics are shown in Table 5. With a fairly low R^2 value in both the train and test sets, our model does not capture much of the variability in the data.

Table 5: Multiple Linear Regression Error Metrics

	Train	Test
RMSE	0.4720575	0.4785082
Rsquared	0.4686585	0.4651986
MAE	0.3474662	0.3504292

The lasso. This technique also yields a linear model of the form seen in multiple linear regression, except lasso attempts to reduce the complexity of the model by performing variables selection to eliminate some predictors (setting their coefficients equal to 0). It does this by adding a penalty term to the typical minimizing RSS formula. Thus, lasso minimizes the quantity

$$RSS + \lambda \sum_{j=1}^{p} \mid \beta_j \mid .$$

The tuning parameter, λ is chosen by the modeler, but finding the best λ is done via cross validation. In this way, lasso and best subset are similar in that they perform variables selection and attempt to shrink the space of predictors used in the model.

In comparison to the multiple linear regression with forward subset selection, we present an ℓ_1-penalized linear regression. We hope the shrinkage of this lasso regression will help identify the parameters for our best model.

Table 6: Lasso Regression Coefficient Values

Variable	Coefficient
(Intercept)	-37.2580145
FIPS_TRACT200	0.0407084
FIPS_TRACT300	0.0141405
FIPS_TRACT800	-0.0003200
FIPS_TRACT1100	0.0238898
FIPS_TRACT1200	0.0637920
FIPS_TRACT14700	-0.0303811
sale_year	0.0211724
BASEMENT_AREA	0.0001153
AGE	0.0017242

Variable	Coefficient
Extra.Fixtures	0.0102340
Full.Baths	0.0544851
OUTDOOR_AREA	0.0001698
LIVING_SQFT_log	0.2560806
LIVING_SQFT_sqrt	0.0040473
PROPERTY_SQFT_log	0.0703155
logZscoreROOF	0.1377892
logZscoreWALL	0.1058283
logZscoreHEAT	0.1128633

As we see in Table 7, R^2 is slightly higher than in our forward selection in Table 5, so the lasso is a marginal improvement. However, for analysis sake, both models perform very similarly - their MAE, Rsquared and RMSE are about the same, except the Lasso regression appears to have a slightly larger Rsquared

Table 7: Lasso Regression Coefficient Values

	Train	Test
RMSE	0.4683169	0.4720037
Rsquared	0.4780337	0.4811342
MAE	0.3458344	0.3483853

Residual plots are available in Figure 4, and indicate relatively normal errors, supporting the assumptions of our model.

2.2 Classification

Logistic Regression. For binary responses (0 or 1), logistic regression calculates the probability

$$P(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots\beta_p X_p}}$$

whose β's are calculated using the Maximum Likelihood function. The classification 0 or 1 is determined by the output probability $P(X)$, with a typical cutoff at $P(X) = 0.5$. The logistic model is linear in X, which can be seen after some manipulation of the above equation. The

log-odds form of the equation is

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

This formulation allows for easy interpretation of the model. For example, increasing X_1 by one unit changes the log-odds by β_1, with all else constant. Distribution of some continuous variables given rent status can be found in the Appendix. Table 3 suggests that the census tract could be a significant predictor for rent status of a property. We will encode rented classification as $1 = \textit{rented}$ and $.0 = \textit{owned}$.

K-Nearest Neighbors (KNN). The KNN classifier algorithm is applied to each test observation x_0. For each x_0 the KNN classifier identifies the k nearest training observations, N_0. It then computes the condition probability for each class j as a fraction of the points whose response is j:

$$P(Y = j \mid X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$

The classification for x_0 is chosen by the highest probability. When k is small, it will be likely that overfitting is occurring because the model is overly flexible, while the opposite is true for KNN when k is large, since the model is near linear for large values of k.

2.3 Hypothesis Testing

When applicable, we can do hypothesis testing to find the statistical significance of each predictor. For Best Subset and Logistic Regression we tested the significance of each β_i with the test:

$$H_0 : \beta_i = 0 \quad H_a : \beta_i \neq 0.$$

Nevertheless, no hypothesis testing was performed for KNN since no coefficients β_i were obtained. As for lasso regression, its hypothesis testing had always been a challenge. Although mathematicians had successfully developed a model named *covariance test statistic* which was suitable linear lasso hypothesis testing and requires only weak assumptions on the predictor *matrix X*, it yet to be a complete model and still requires further improvement(Lockhart et al. 2014). Hence, this study did not perform hypothesis testing for linear lasso model or logistic lasso model.

3 Results

3.1 Logistic Regression Results

Table 8: Summary of FIPS Logistic Regression results

	Owned	Rented
Predicted Owned	51	34
Predicted Rented	1076	3526

Table 9: Summary of AGE Logistic Regression results

	Owned	Rented
Predicted Owned	0	0
Predicted Rented	1127	3560

Two logistic regression models were fitted individually with FIPS_TRACT and AGE as predictors for classifying housing rental status. Each model was trained with 70% of the data (classData) and validated with the remaining 30%. Both logistic models showed high false positive rates meaning that neither model is suitable when a property is owned (see Table 8 and Table 9). In fact, for the logistic model just using age as a predictor, the model never predicts a property as owned, as seen in Table 9.

Nevertheless, ($\beta_{AGE} \approx -0.0083$) was a significant predictor for rental status with $p < 0.05$, increasing AGE by one year changes the log odds by approximately -0.0083.

The model using just the FIPS_TRACT predictor rental status gave slightly better results. The significant census tracts are shown in Table 10. All β's had corresponding $p < 0.001$. We can see from Table 8 that the model predicted some owned properties correctly, but still, performance is poor as the confusion matrix shows. However, some census tracts were statistically significant ($p < 0.05$). The results are summarized in Table 10.

Table 10: Summary of significant census tracts coefficients

FIPS_TRACT	β value
FIPS_TRACT200	$\beta = -0.5918$
FIPS_TRACT300	$\beta = -0.8587$
FIPS_TRACT400	$\beta = -0.5952$

FIPS_TRACT	β value
FIPS_TRACT500	$\beta = -0.6139$
FIPS_TRACT900	$\beta = -0.7198$
FIPS_TRACT1000	$\beta = -0.9122$
FIPS_TRACT1100	$\beta = -2.0874$
FIPS_TRACT1200	$\beta = -1.8403$
FIPS_TRACT1400	$\beta = -1.1661$
FIPS_TRACT11805	$\beta = -1.5364$
FIPS_TRACT13202	$\beta = -2.7460$
FIPS_TRACT13301	$\beta = -1.4428$
FIPS_TRACT13501	$\beta = -3.5513$
FIPS_TRACT13503	$\beta = -2.5704$
FIPS_TRACT14700	$\beta = -1.0543$

Overall, increasing either age or FIPS_TRACT will result in decreasing the probability of a house being classified as rented, but the effect is nt strong given that some β's are very close to 0.

3.2 KNN Results

Table 11: Summary of KNN test results

	Owned	Rented
Predicted Owned	0	0
Predicted Rented	1127	3560

A KNN model was fitted to predict the classification of rent status with three predictors: FIPS_TRACT, OUTDOOR_AREA, and Number.of.Bedrooms. The model was trained with 70% of the data and validated with test data from the remaining 30% of the data. The validation results for KNN with $k = 10$ are shown in Table 11. This model has an accuracy rate of 0.7595, which is slightly higher than the ratio of rented properties to total properties(approximately 0.7461). Along with the high false positive rate, this K Nearest Neighbor model is not good at predicting when a property is owned. One can see from Figure 3 that the curve indicates our model does slightly better than no class separation whatsoever. That is, the AUC, area under the curve, is greater than 0.5 but not very large, which reflects the model's ability to accurately

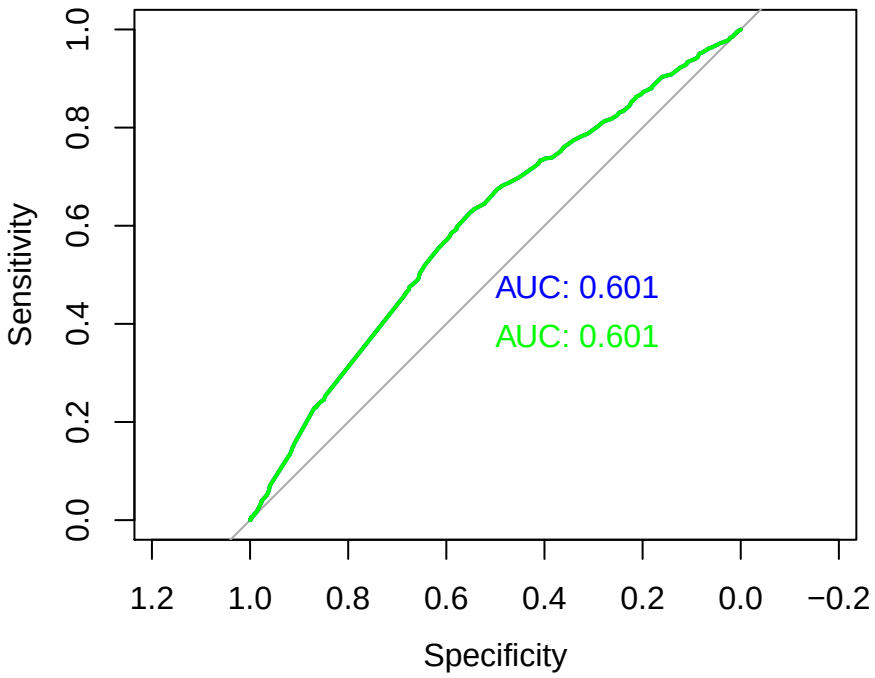


Figure 3: Sensitivity and specificity of KNN classifier

predict some, but not most, false negatives.

Because the K Nearest Neighbor does not allow for estimation, we are unable to discern which predictors most impacted the model's performance. Obviously demographic predictors like race, ethnicity, age, marital status are related to rent status as census data consistently shows (Office of Policy Development and Research (2017)), but our model did not use such predictors. Our results from K Nearest Neighbor classification that predictors that are intrinsic to the property itself—physical features— are not adequate for accurately predicting whether the property is rented or owned.

4 Conclusion

Our results on the classification question—can one predict with sufficient accuracy the rent status of a home in Lancaster City using intrinsic characteristics of the house—do not show strong support in the affirmative. One of our statistical models was unable to accurately assign a house as owned, and the other model only did slightly better. That model still incorrectly assigned the property the status of rented even though in reality it is owned at a high rate. If a researcher or investor is interested in predicting the rent status of an individual property or a collection of properties in a neighborhood, it is advisable that they use other predictors than

the ones used in our models, such as the age of the home, the number of beds, and the square footage. Instead, characteristics about the land owner or resident, following prior research, would most likely yield better results for Lancaster City.

5 Appendix

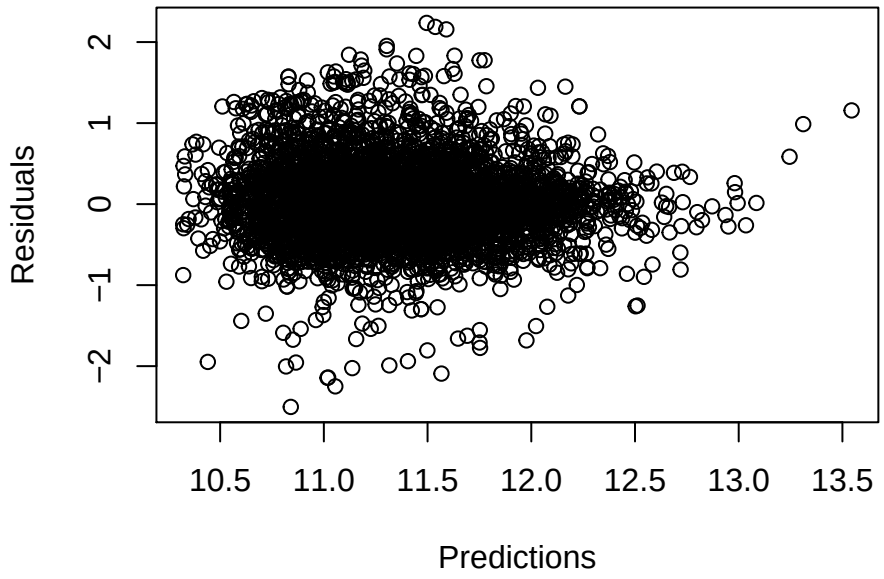
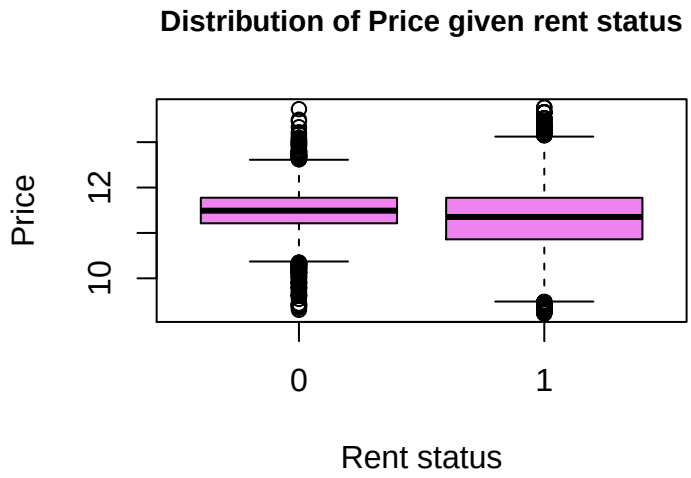
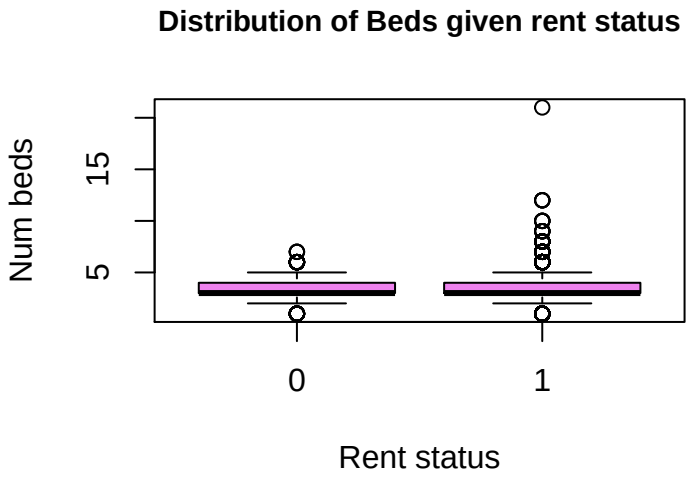


Figure 4: Error Plot of Lasso Regression





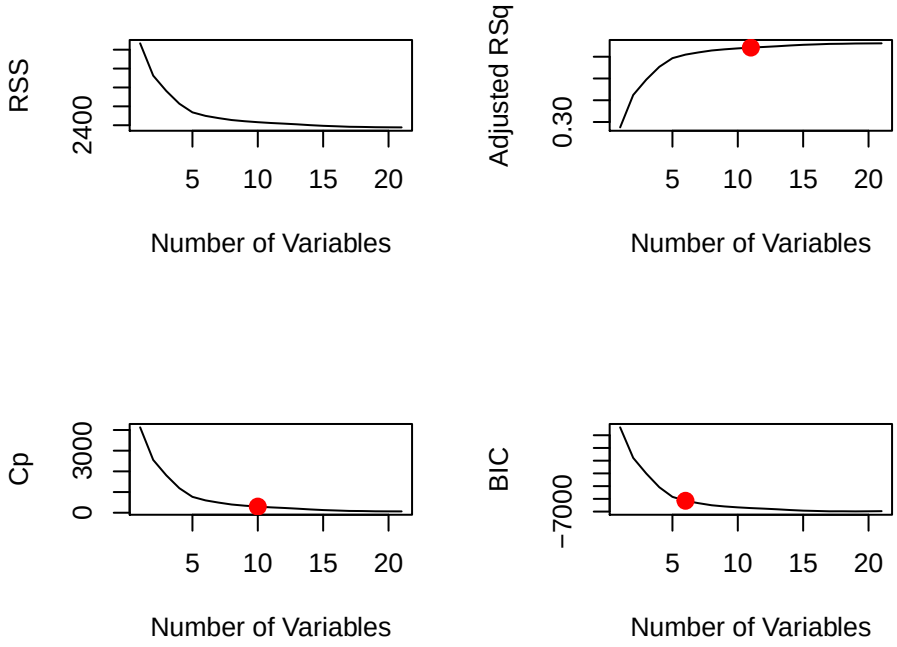


Figure 5: Forward Subset Error for Number of Parameters

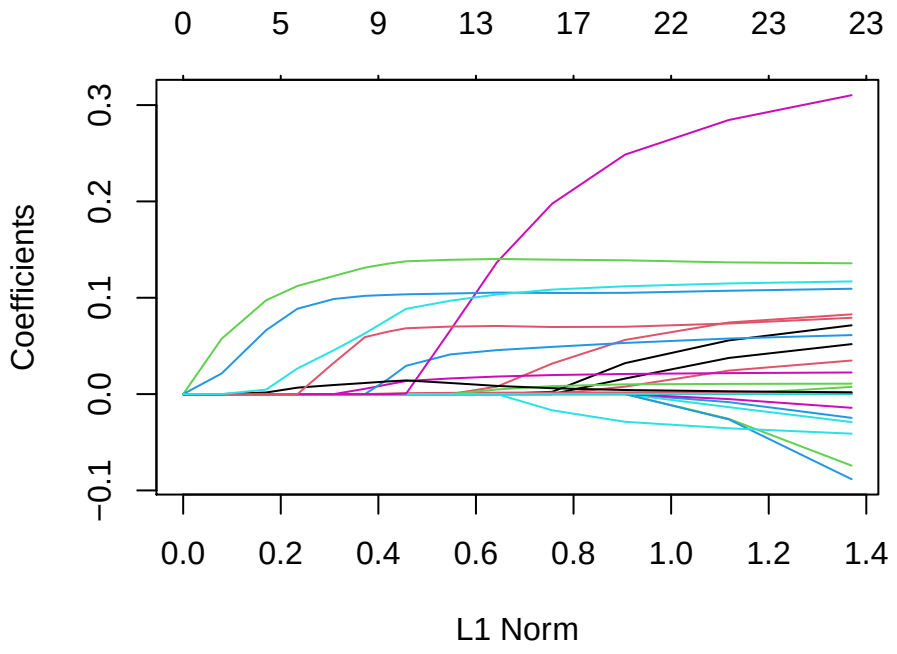


Figure 6: Linear Regression Lasso Coefficients

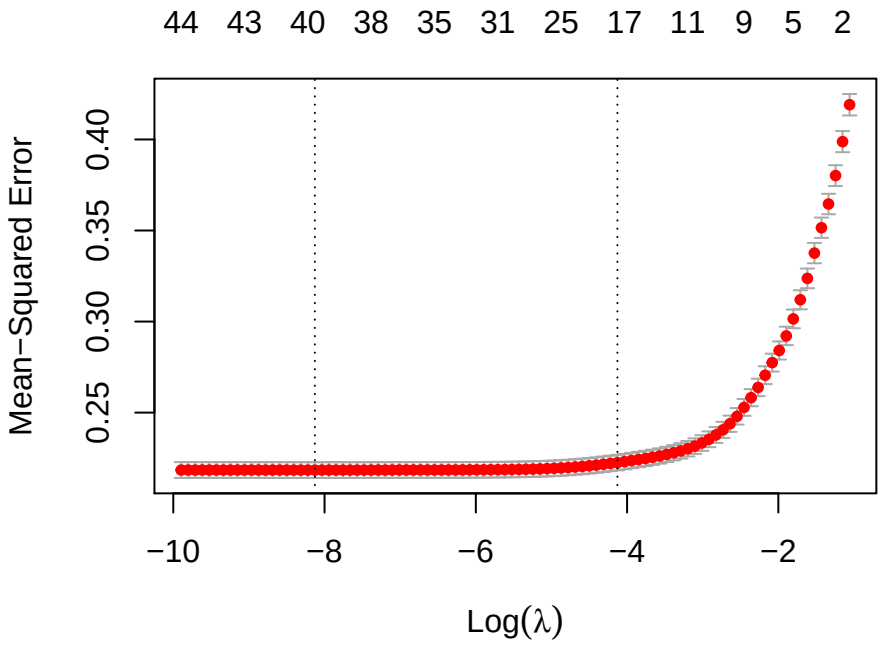
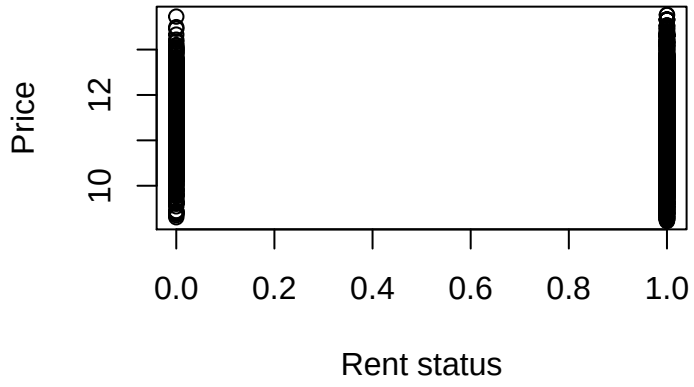


Figure 7: Linear Regression Lasso MSE


Distribution of Price given rent status
Price
10
12
0.0
0.2
0.4
0.6
0.8
1.0
Rent status

6 References

Krishna, Priya. 2019. "A Global Feast in an Unlikely Spot: Lancaster, Pa." *The New York Times*. https://www.nytimes.com/2019/07/23/dining/lancaster-pennsylvania-restaurants.html.

Lockhart, Richard, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. 2014. "A Significance Test for the Lasso." *Annals of Statistics* 42 (2): 413.

Office of Policy Development and Research. 2017. "American Housing Survey: Renters Status." Department of Housing and Urvan Development. https://www2.census.gov/programs-surveys/ahs/2017/infographs/2017/%20Housing/%20Profile/%20Renters/%20Profile.pdf.

Strasser, Franz. 2017. "Lancaster, Pennsylvania: America's Refugee Capital." *BBC News*. https://www.bbc.com/news/av/world-us-canada-38776233.