Using Computer Vision and Machine Learning to Unlock Historical Data

Jun Tao Luo

CMU-CS-23-111

April 2023

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Matthew Gormley, Chair
Rayid Ghani

Submitted in partial fulfillment of the requirements for the degree of Master of Science.

Keywords: Machine Learning, Computer Vision, OCR, Historical Records

For my late grandfather who instilled a love of learning in all of us.

Abstract

Historical administrative records (e.g., property transfers, birth certificates, census data) can be extremely valuable for academic research and industry applications. However, such data is rarely digitized or accessible in analyzable formats.

We demonstrate how machine learning and computer vision methods can be combined to create a cost-effective digitization technique for historical property tax assessment records. We show how image processing and optical character recognition (OCR) deep learning models retrieve records with a mean absolute percentage error (MAPE) of 14.72%. For cases where OCR cannot be applied, such as when scanned documents are not available, we combine a small sample of manually labeled historical data with contemporary feature data to build regression models that retrieve records with a reduced accuracy of 17.48% MAPE. Both methods present a substantial saving over manually digitizing the same data, with OCR achieving a cost reduction of 78% and the regression model achieving a cost reduction of 89%.

Acknowledgments

I would like to thank my thesis committee Matt Gormley, Junia Howell and Rayid Ghani for giving me this opportunity to work on this exciting subject and giving me a wealth of advice along the way. Without their valuable insights and advice, this work would not have been possible. I also want to thank Mihir Bhaskar for working with me on this project and his collaboration on the data cleaning and regression models. Finally, I would also like to thank my family and friends who supported me along the way.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

From trade records for economic research [11] to death certificates for epidemiological studies [2] to church records for demographics analysis [29], historical administrative data has wide applications in academic research and industry applications. Often, this data is stored in physical formats with handwritten information that is not easily accessible or usable by data analysts and scientists in its original form. This is especially challenging for data recorded in tables and forms where the structure of the document, such as the relationship between rows and columns, is critical to correctly parsing the document. Current methods for obtaining usable information from these historical records involves manually entering data from scanned or paper documents, see Figure 1.1. This process is prohibitively costly for large projects.

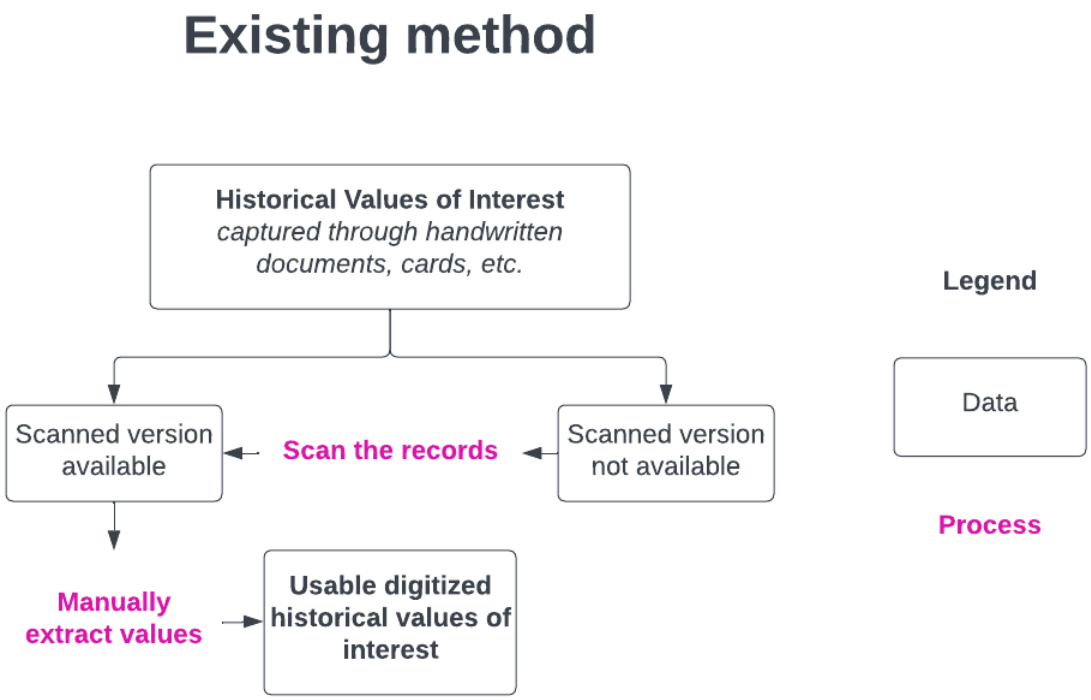


Figure 1.1: Current Methodology for Digitizing Records

Our focus is on administrative historical data recorded in relatively consistent formats. Examples of such documents include decennial censuses, port of entry records, birth and death certificates, and property permits, deeds, and tax assessments. Given that the preservation of historical records is inconsistent, with some records getting damaged or lost over time, our goal is to develop a holistic approach with a menu of options for digitization. This enables scholars and practitioners to weigh the various cost and accuracy trade-offs and select the best approaches for their specific project.

To test and evaluate this approach, we use historical property appraisal cards from Hamilton County (Cincinnati) and Franklin County (Columbus). Historical property appraisals are of particular interest for equitable housing research which explores the connection between race and property values [14]. In this field of research, housing equity scholars have proposed fairer al-

ternative appraisal systems which value properties using data on the actual costs of construction rather than comparable sales [13]. This system is untested because it relies on digitized construction cost data, and in Hamilton County, approximately 55% of all buildings were constructed before such data exists (1960). Researchers argue that early historical property appraisals are a good proxy for the original construction costs of these old buildings because of appraising practices at the time.

The goal of this case study is to digitize the earliest available building appraisal value from the historical property cards, such that it can serve as a usable measure of original construction costs for buildings of this vintage. We propose two distinct methods for obtaining estimates that differ greatly in their cost and accuracy: (1) automatic extraction of the estimate from scanned appraisal cards using OCR and (2) regression models capable of estimating the 1930s value from contemporary data about the parcel. These approaches are illustrated in Figure 1.2. For parcels where scanned appraisal cards are available, We use computer vision techniques and optical character recognition (OCR) models to extract the building appraisal value from scanned property ownership cards in 1933, which is the earliest year available. For cases where scanned appraisal cards are not available, we combine manually labeled ownership cards with contemporary information about the land parcels[1] and building characteristics to construct a training set. We then build regression models to estimate the original construction cost using only the contemporary data. These two methods are also mutually beneficial. OCR results can also be used as additional training labels for regression models and in cases where OCR methods cannot be applied, the regression model can be used to provide an estimate.

[1]Parcels are the main administrative unit for properties

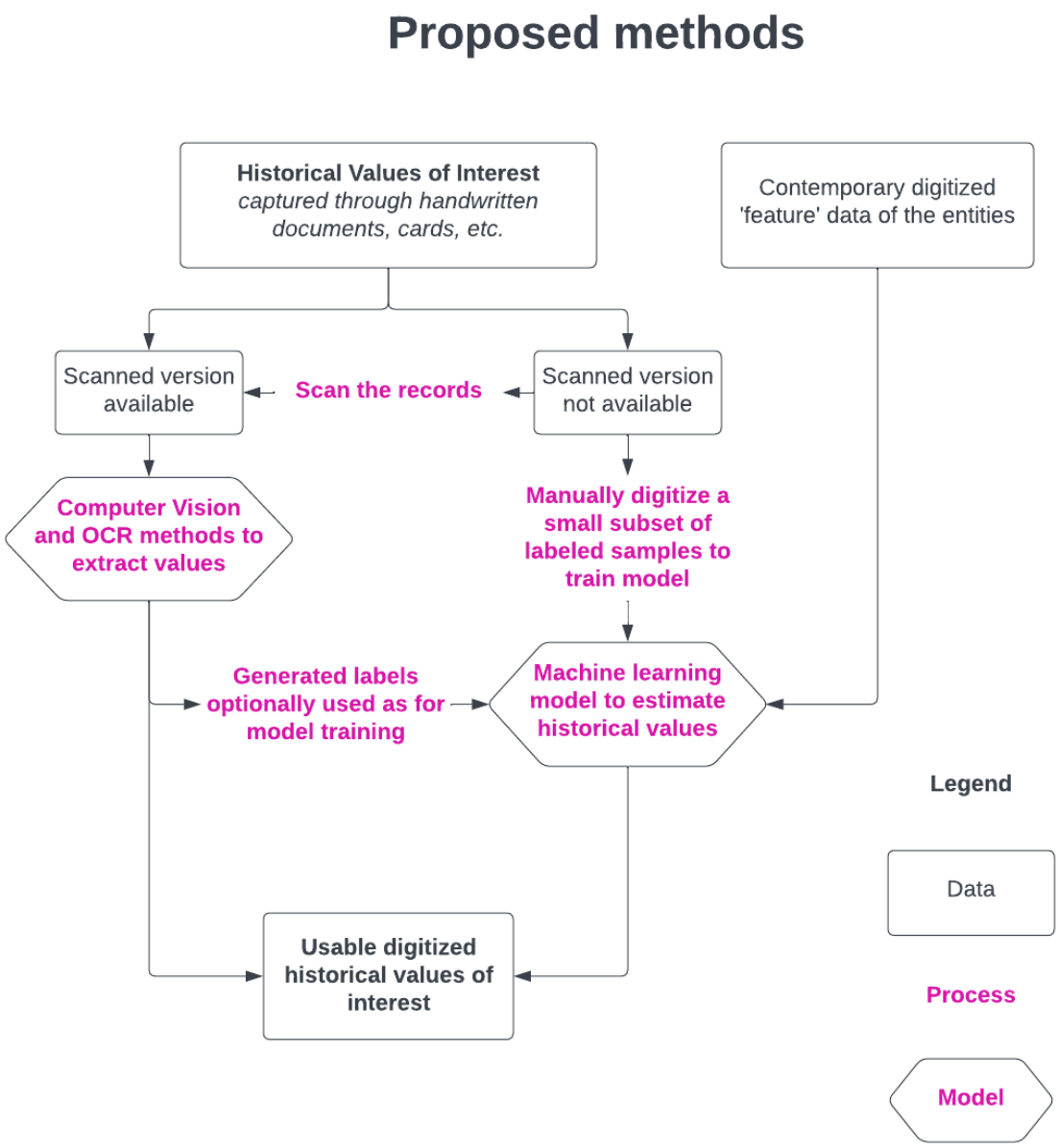


Figure 1.2: Proposed Methodology for Digitizing Records

1.1 Related Work

Recent interests in digitizing historical documents have used OCR technologies to extract data from scanned documents including balance sheets [6], and newspapers [4] [19]. Although some of this work has focused on numeric extraction from historical censuses [21] and church records [29], it does not address the challenge of semantic understanding of tabular documents. This involves using the surrounding context such as table borders and relative positions of text segments to divide the document into rows and columns, similar to approaches in described in tabular OCR works [22], [10], [24]. We also incorporate techniques from previous OCR works including TesseractOCR [26] and transformer based TrOCR [17] in developing our models.

Similar to our use of regression models for historical records, we find that machine learning models have been used to link families and individuals across historical census records, but not estimate specific values from these records [9], [23]. In the realm of real estate, there is a rich literature using machine learning to predict contemporary property value and sale price given its industry applications to real estate valuation and transactions [30], [12], [28], [3], [27].

Chapter 2

Experimental Setting

This chapter details how we translate the broad approach defined in Figure 1.2 to our use case. We use Hamilton County (Cincinnati), Ohio as a starting point because of the public availability of both scanned historical appraisal cards and contemporary property information. First we describe the available data, processing needed to apply our methods, and how we arrived at a subset of parcels for analysis (§2.1, 2.2). We then propose a simple baseline method of estimating building construction cost to benchmark our results (§2.3).

2.1 Data Sources

We obtain administrative data made publicly available by the Hamilton County Auditor on their website. [1] The records listed below are linked by a unique identifier at the land parcel level. [2]

[1]The Auditor is the County's Chief Fiscal Officer and Property Assessor. Their website is: https://hamiltoncountyauditor.org/

[2]We use "parcels" and "properties" interchangeably, with "buildings" used to refer to human made enhancements to the land.

Target Data: Historical Property Ownership Cards

These documents are scanned images of the historical property details for a parcel. This includes ownership and transfer information as well as land and building valuations. While dates are missing for most of the valuations, process documents suggest that the assessments that generated these values followed the same three year cycle used today, with the first values generated in 1933. An example of this document can be seen in Appendix A. Our main target for digitization comes from this body of documents: the initial value of the building, as recorded in 1933, which serves as a proxy for its original construction cost. The distribution of this variable, based on 10,452 randomly selected hand-labeled samples, is shown in Figure 2.1.

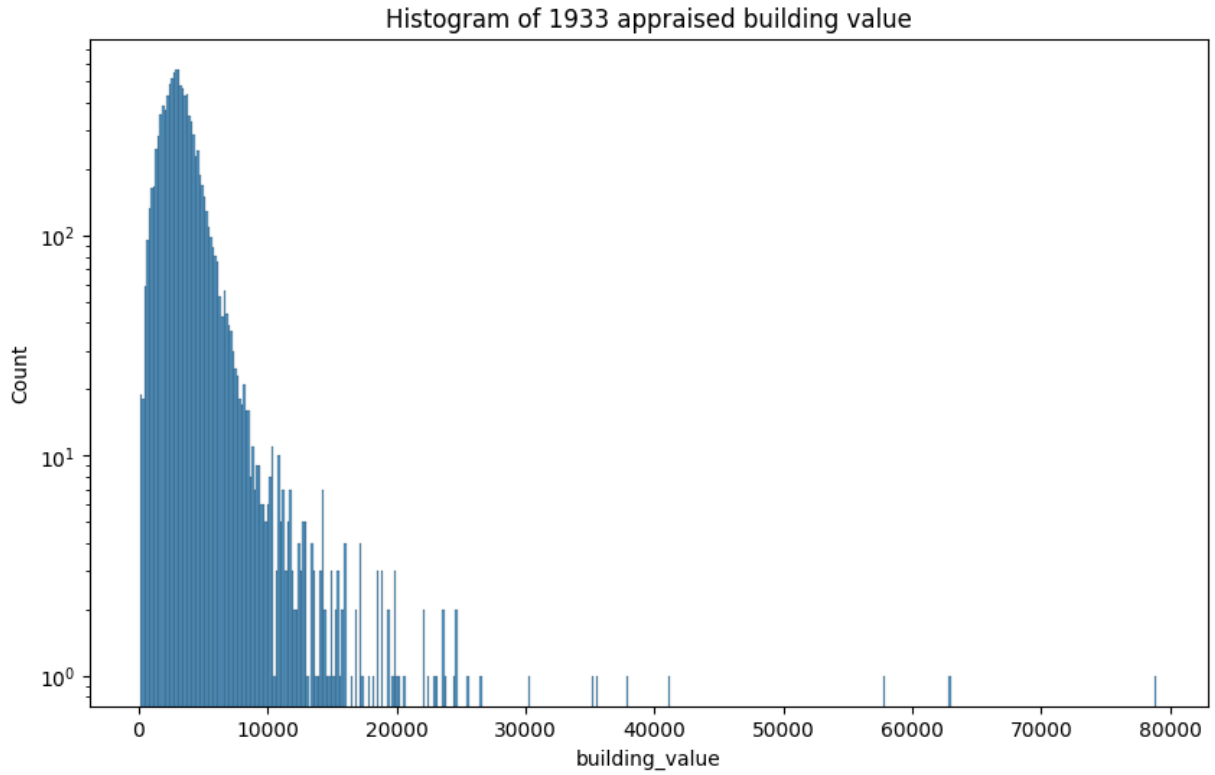


Figure 2.1: Histogram of Target Value

Table 2.1: Features built from contemporary data

Square footage	attic, basement, floor 1, floor 2, half-floor, total livable area
Building characteristics	stories, style, grade/condition of building, exterior wall type, basement type, heating type, air conditioning type, total number of rooms, total full and half bathrooms, number of fireplaces, garage type and capacity
Parcel characteristics	land use code, neighborhood, number of subparcels

Feature Data: Contemporary Tax Assessment Information

Every three years, the County Auditor updates all property assessments for tax purposes [1]. We use data from the latest assessment, in 2020, as the most updated and comprehensive set of parcels. The data includes information about the parcel's administrative status, information about the physical characteristics of any buildings on the parcel, and information about valuations and sales. A full set of features[3] used from this data is listed in Table 2.1.

2.2 Data Processing

The raw data for contemporary tax assessment information could be downloaded directly as structured Excel files from the source. Several data cleaning steps were needed to process the data into a format ready for analysis, including handling nonsensical values, grouping categories, and creating consistency in formats across source tables. These are detailed in Appendix D.

The 353,973 parcels found in the contemporary data were subset based on the following criteria:

[3]We use one-hot encoding for categorical features

1. **Parcels defined as residential**: building characteristics such as rooms and bathrooms are not captured for commercial buildings
2. **Parcels with one single, finished building**: the data does not contain a building identifier, making it impossible to know which building the characteristics pertain to for parcels with multiple buildings. Hence, we focus on only parcels with one construction.
3. **The building was constructed before 1930**: since the target value of interest is the appraised building value recorded in 1933, we only consider parcels that had a construction before 1930.
4. **No data inconsistencies between tables**: to ensure all the feature information could be used, we ignored parcels that did not match across source tables or contained inconsistent information about building characteristics between source tables.

The resulting set of 59,378 parcels forms the overall sample of interest for Hamilton County. Of this set, we were only able to successfully retrieve 56,037 scanned documents, which indicates 5.6% of the parcels are missing their ownership card documents. To ensure we do not introduce bias due to these missing documents, we perform a Classifier 2 Sample Test which is detailed in Appendix B. Next, we perform basic pre-processing on the documents including cropping, rotating, and conversion to grayscale.

In order to create labels for the regression models, we randomly sampled 12,423 of the retrieved ownership cards for manual labeling. See Appendix C for additional details about the manual labeling process. For a breakdown of the number of samples after each processing step, see Figure 2.2.

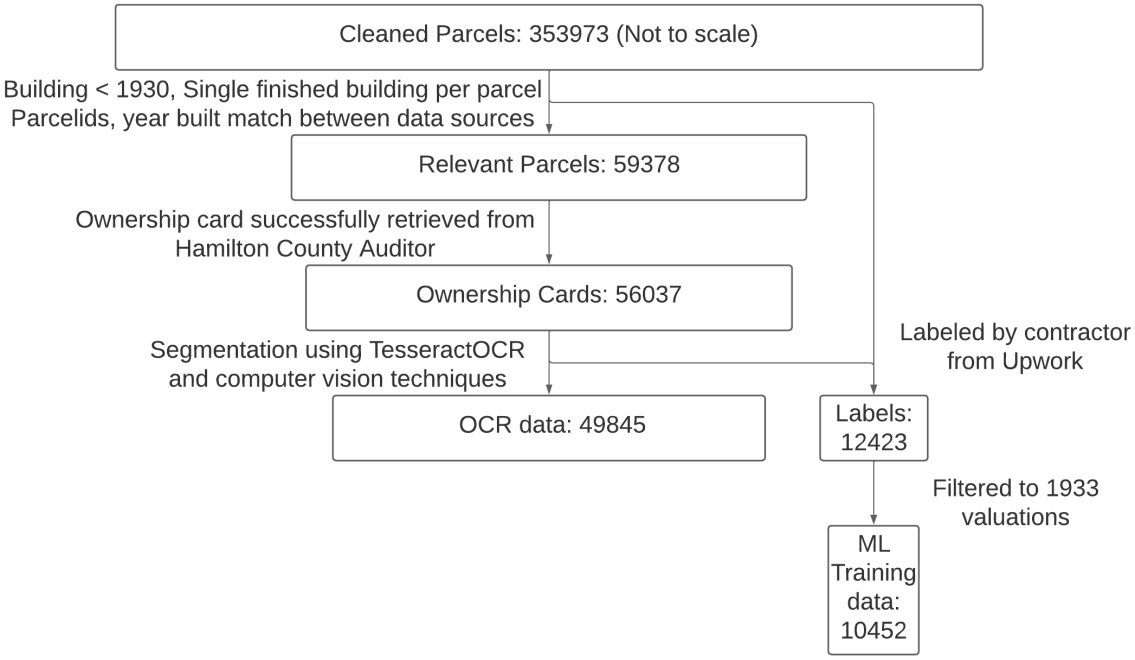


Figure 2.2: Data Processing Flow

2.3 Baseline Model

Our baseline model was created with the following question in mind: in the absence of machine learning, what would be required for an investigator to estimate 1933 tax assessment?

We use a national model for estimating construction costs based on the American Housing Survey, that takes into consideration region and urban/rural classification. Specifically, we use the following equation for Midwest urban regions which contains Hamilton County.

$$
\begin{aligned}
&\text{Estimate Construction Cost} \\
&= -231522.85458 + 116.894831 * \text{Year} \\
&+ 1222.49492 * \text{Basement} + 6267.1243 * \text{Heating} \\
&+ 1769.41924 * \text{Central AC} + 1579.6329 * \text{Total rooms} \\
&+ 991.89863 * \text{Bathrooms} - 4.9088314 * \text{Half Bathrooms} \\
&+ 2373.59715 * \text{Garage}
\end{aligned} \tag{2.1}
$$

In Equation 2.1, the variables are defined as follows:

- Year - The year the home was built
- Basement - A dichotomous variable indicating the presence of any basement
- Heat - A dichotomous variable denoting the presence of heating system
- Central AC - A dichotomous variable denoting the presence of central air conditioning
- Total Rooms - The total number of rooms in the building
- Bathrooms - The number of bathrooms in the building
- Half Bathrooms - The number of half bathrooms in the building
- Garage - The number of cars the garage can hold

Chapter 3

Methods

This chapter described the details regarding the models and experiments that were performed as part of this work.

Section 3.1 describes the computer vision and OCR workflow: that is, given the availability of scanned cards, the methodology we use to extract the earliest appraised value. Section 3.2 describes how we build the machine learning models to estimate the value using contemporary feature data. Section 3.3 details how we combine these two workflows by incorporating predictions from OCR into the training data for the regression model.

Finally, Section 3.4 describes an experiment to test how well the model generalizes to Franklin County (Columbus), Ohio. A model that could generalize across cities and regions would unlock significantly more historical appraisal data for use by researchers and practitioners at lower cost than building city-specific models.

3.1 Computer Vision and OCR

3.1.1 Tabular Data Segmentation

Although there are many solutions for OCR, we find that no existing method was able to accurately extract values from the tables in our documents. The first challenge is to recognize the tabular structure of the ownership card documents and locate the relevant information. Given that we use the building value of the first recorded appraisal in the Hamilton ownership documents as a proxy for our target variable, this involves obtaining the first entry of the "BUILDINGS" column. To accomplish this, we use a customized process for segmentation which involves using TesseractOCR to locate the column header "BUILDINGS" then using Hough Transform to locate surrounding row and column divisions for cropping individual table cells. For more details about the segmentation step, see Appendix F. The individual cropped cells images are then passed to a higher accuracy system for OCR called TrOCR.

3.1.2 Optical Character Recognition (OCR) Models

Given that our task involves recognizing only numerical values, it is challenging to use off the shelf OCR solutions or pretrained models such as TesseractOCR or TrOCR since these models predict all characters, including digits, punctuation and letters, and perform poorly on our dataset. Initial experiments show that these pretrained models would often confuse letters and digits including recognizing the digit 0 with the letter O and the digit 1 with lowercase L or uppercase I. To address this type of error, we perform additional find tuning using a mixture of different datasets including CAR-B (handwritten digit strings from scanned checks) [7] and DIDA (historical handwritten digit dataset) [16]. For detailed descriptions of our OCR experiments, see G.

Table 3.1: Chosen Regression model

Model class	Random forest regressor
Number of estimators	2500
Max depth	200
Minimum samples for split	4
Max features	sqrt

3.2 Regression Model

We formulate the task of predicting a historical value from contemporary property data as a standard regression problem where the target value to be predicted is the labeled 1933 building appraisal value. As mentioned in Section 2.2, we have 10,452 parcels with labels collected by hand, which we merge with the contemporary feature data outlined in Table 2.1 to create the training and test matrices. We use an 80-20 train-test split, and employ 5-fold cross validation within the training set for hyperparameter tuning.

We use a stepwise approach to model selection. First, using a single set of default hyperparameters, we train many different model classes and observe performance on a validation set. The results of this are in Appendix H. We then select the best performing model classes, and conduct a more extensive hyperparameter grid search, selecting the best model using the 5-fold cross validation root mean squared error (RMSE).

This approach leads us to choose a random forest regressor as the best model. The hyperparameters of this model are in Table 3.1.

3.3 Augmented Regression Models

While OCR methods and regression methods are two separate approaches for predicting the same target variable, they accomplish their task using different inputs and techniques. These

two methods are complementary to each other in that they can be combined in various ways to improve performance. In this work, we use the trained OCR model to create annotated labels for training the regression model. This allows the use of all 56,037 retrieved scanned documents for training and testing of the regression model instead of only the 12,423 manually labeled samples. We show in Section 4.3 that this improves the performance of the regression models.

3.4 Model generalization

3.4.1 Generalization to Franklin County

To test whether our regression model generalizes to a different city, we collect test data from Franklin County (Columbus), Ohio. Similar to Hamilton County, Franklin County has publicly available contemporary property appraisal data[1]. Since the appraisal cycles of both counties did not exactly align, our target variable is the closest appraisal year to 1933 that we could find data in the historical cards. For 99% of cards in our sample, this is the 1931 appraised value.

Using the same logic as in Hamilton, we subset the universe of parcels in Franklin to a sample of 42,100 parcels that have one residential building built before 1930. We manually hand-label a randomly drawn subset which provides us with a small test set of 506 observations.

To apply the trained model to make predictions in Franklin County, we had to ensure that the features in Franklin County were comparable to those used in the Hamilton model. While some of the important features were common (e.g., square footage of floor 1), there were several features not available in Franklin County or captured in a different format (e.g. presence of attic captured rather than specific square footage). To test generalization, we train the model with only the subset of features that were comparable across both counties, and use this limited model to report performance on the Franklin County test set. See Appendix E for more details on the

[1]Made available by the Franklin County auditor `https://apps.franklincountyauditor.com/Outside_User_Files/2023/` accessed on 2023-03-15

feature subset used.

3.4.2 Generalization of Regression Models to OCR Failures

As another dataset for which we want to test our regression model's generalization, we collected the group of 6,192 Hamilton County parcels for which our OCR methods failed during segmentation. We manually labeled a random sample of 778 of these cases and evaluate our augmented regression model's predictions on these samples and determine whether the model's performance on this subset is similar to those from the augmented regression experiments.

See Appendix E for more details on the feature subset used.

Chapter 4

Results

In this section, we present the results of the experiments using the methods described in the previous section. We start by examining the performance of the baseline model (§4.1). Then the results of the OCR methods (§4.2.1, 4.2.2) and regression models (§4.2.3) are discussed. Finally, we discuss the results of the regression model augmentation and generalization experiments (§4.3, §4.4)

The statistics of the best performing models from our experiments are shown in Table 4.1. We sampled 20% of the manually generated labels and filtered the samples to values within the 5-95 percentile range of the overall target value distribution to remove outliers.[1] The resulting test set is used to report the metrics in this section. We use several common statistical evaluation metrics for regression tasks including coefficient of determination (R^2), Mean Absolute Error (MAE) which in this case measures the amount in US Dollars, Mean Absolute Percentage Error (MAPE), Root Mean Squared Percentage Error (RMSPE), the Median Percentage Error (MPE), and the percentage of test cases where we predicted a value that is within 5%, 10%, or 20% of the true value. We report results on buildings in the middle 90% of properties based on appraised value (i.e., 5th to 95th percentile), to reduce the effect of outliers.

[1]The generalization model reports performance on the smaller test set of 506 observations in Franklin County.

Table 4.1: Prediction performance of evaluated models

Metrics	Baseline	OCR	Regression	Augmented	Generalization [1]
R^2 (higher is better)	0.0195	0.6264	0.6177	0.7428	0.3846
MAE (lower is better)	\$907	\$492	\$489	\$452	\$571
MAPE (lower is better)	33.89%	14.72%	17.48%	16.12%	22.72%
RMSPE (lower is better)	52.10%	40.04%	27.73%	24.01%	38.71%
MPE (lower is better)	21.49%	0%	10.60%	11.27%	28.31%
Within 5% of True Value (higher is better)	11.84%	85.36%	25.81%	24.39%	15.70%
Within 10% of True Value (higher is better)	25.44%	85.39%	48.06%	45.85%	31.40%
Within 20% of True Value (higher is better)	47.32%	85.40%	73.44%	74.55%	62.50%

4.1 Baseline Model

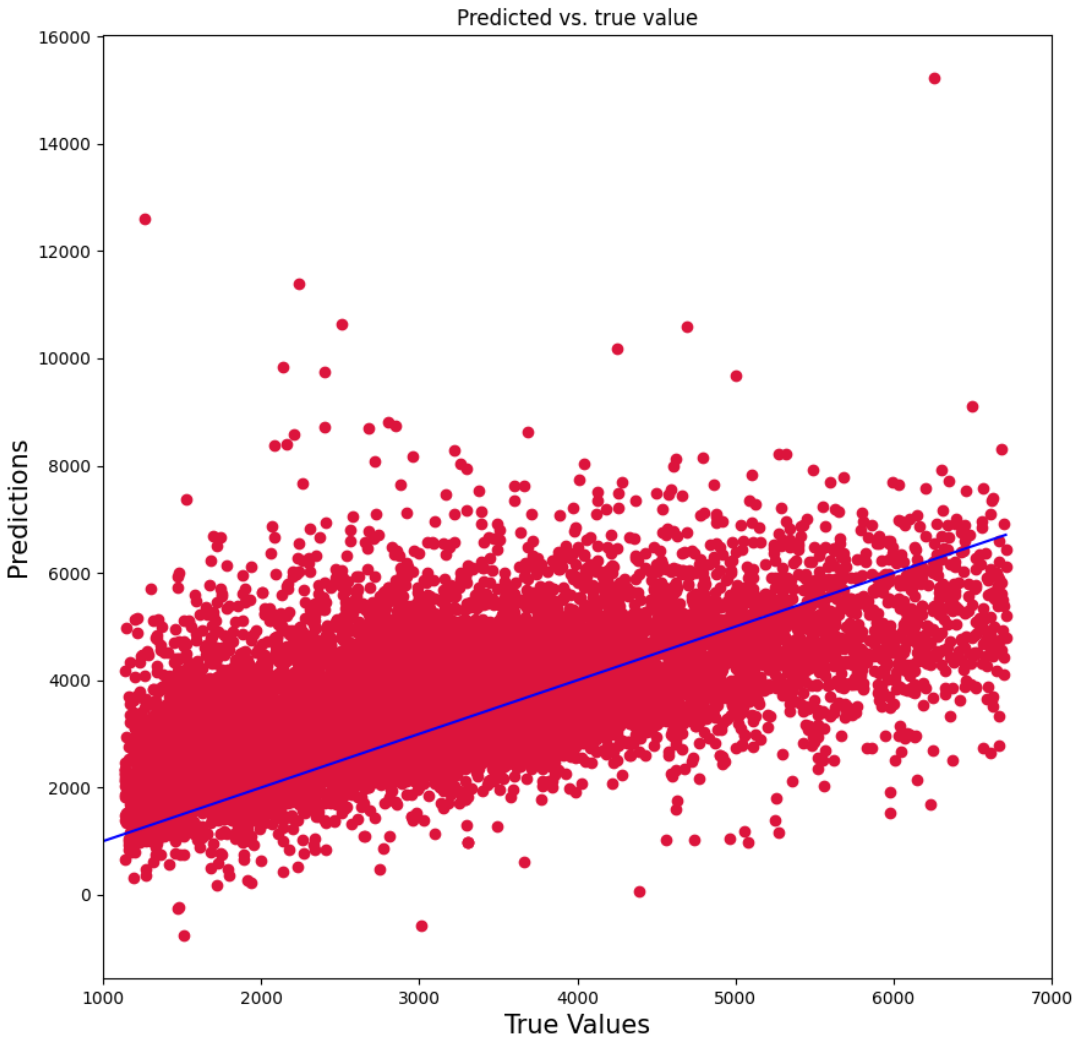


Figure 4.1: Baseline Model Predictions

As a benchmark to compare our proposed methods against, we measure the performance of the predictive power of the baseline model. We use the data for Hamilton county parcels and compare the estimates generated by Equation 2.1 with the hand labeled ground truth values. The

predictions of this model are shown in Figure 4.1. Analysing the errors between the predictions and true values shows that the model makes large errors in its prediction, as illustrated by large MAPE, RMSPE and MPE values, and these errors are common with only 11.84% of the predictions falling within 5% of the true values. This model performs much worse than our other approaches since it is developed using regional level data (e.g. Midwest Urban) but does not account for trends existing at the county level like the other methods proposed in this work. The baseline model also does not include square footage features, which are shown to be important for our regression models (§4.2.3). The fact that such a simple model performs this well is unexpected.

4.2 Computer Vision and OCR

In this section, we present the the results from our segmentation and OCR experiments which are applicable to scenarios where scanned historical documents are available for direct extraction of target values.

4.2.1 Tabular Data Segmentation

We find that existing solutions that perform segmentation such as TesseractOCR performs poorly on our tabular data, see Appendix F and **??** for details, so we developed a custom segmentation method using Hough Transform and evaluate its performance on our data set. Since this component does not directly produce estimates for the construction cost of the building it is not included in Table 4.1 even though it contributes to the performance of the OCR model. There are two metrics of interest when evaluating the segmentation method: the success rate of extracting a segment and the accuracy of extracting the correct segment. For the success rate, we use our segmentation algorithm on 56,037 documents and are able to successfully extract segments for 49,845 of them giving a success rate of 89.0%. To evaluate the accuracy, we randomly sample

499 ownership documents and examine the tables to compare if the extracted table segment is correct. We find only 1 error case where the segment represents the second cell in the column instead of the first, giving an accuracy of 99.8%.

4.2.2 OCR Models

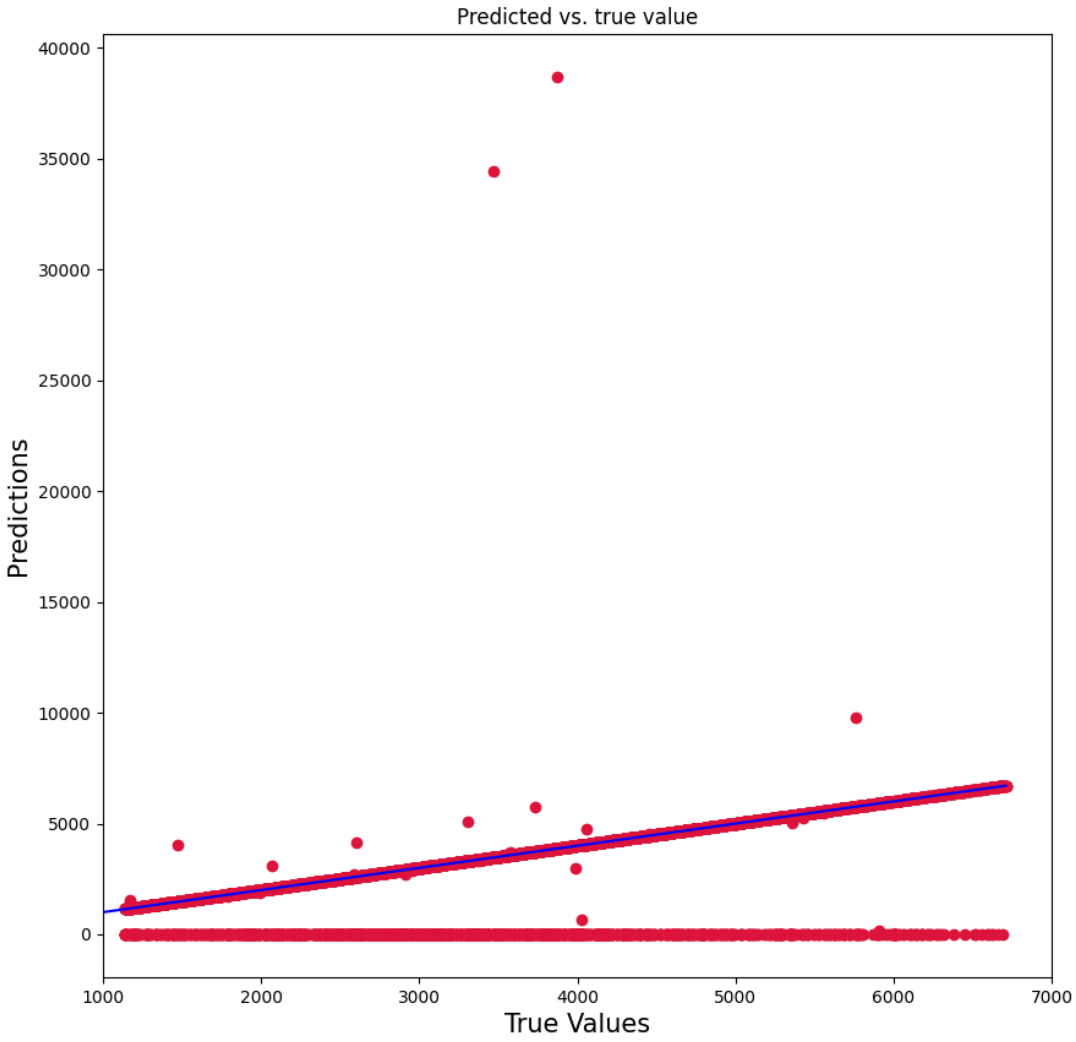


Figure 4.2: OCR Model Predictions

Table 4.2: Prediction metrics of OCR models for different confidence thresholds

Metrics	Top 90%	Top 95%	Top 99%	All 100%
R^2 (higher is better)	0.7663	0.6958	0.6350	0.6264
MAPE (lower is better)	5.417%	10.36%	13.86%	14.72%
RMSPE (lower is better)	26.21%	34.26%	38.97%	40.04%
MPE (lower is better)	0%	0%	0%	0%
Within 5% of True Value (higher is better)	94.68%	89.73%	84.19%	85.37%
Within 10% of True Value (higher is better)	94.71%	89.76%	86.25%	85.39%
Within 20% of True Value (higher is better)	94.72%	89.77%	86.26%	85.40%

We find the best performing OCR model to be TrOCR fine tuned on a mixture of our Hamilton county dataset combined with additional handwritten digit data from CAR-B. Fine tuning on the DIDA dataset was found to be detrimental since the digit strings are primarily year values recorded in church documents caused the fine tuned TrOCR model to incorrectly predict values between 1800-1940 more often. The results from the best performing TrOCR model trained on 7375 entries randomly sampled from our Hamilton county dataset and 3000 entries from CAR-B, see Figure 4.2. We find that this model is relatively accurate with low MAPE and MPE values. Upon further analysis, we find that while errors are rare, as evident by the fact that 85.36% of all predictions falling within 5% of their true values, the magnitude of the errors are large. This is often due to the insertion or deletion of digits which creates extremely large errors and results in large RMSPE and is reflected by the outliers in Figure 4.2. We note that another common error case is where TrOCR fails to detect recognizable digits. In this case, it will output a blank prediction which is converted to a prediction value of ”0” for the purpose of our analysis. Fortunately, these errors are usually accompanied by a low confidence score which allows these low confidence predictions to be filtered. By choosing an appropriate threshold, we can achieve an exact match accuracy of up to 99.4%. To evaluate the impact of this filtering on the outputs of the model, we report the accuracy metrics for retaining top 90%, 95% and 99% of the most confident predictions, see Table 4.2. We see significant improvements in the model performance if we retain only the top 90% of the most confident predictions, achieving and MAPE of 5.417% and able to make a prediction within 5% of the true value for 94.68% of the test cases.

4.2.3 Regression Models

In this section, we present the results of our regression model which is used for target value estimation when scans of the historical documents are not directly available. The chosen random forest regressor model predicts the target value with an MAPE of 17.48%, which is a substantial improvement over the baseline method. As seen in Figure 4.3, the regression model seem to

perform worse on higher-value properties, with larger over-predictions and under-predictions. Many of the square footage-related features and other building characteristics such as grade, wall type, and number of rooms are in the top 10 most important features based on impurity-reduction. See Appendix I for a plot of the feature importances.

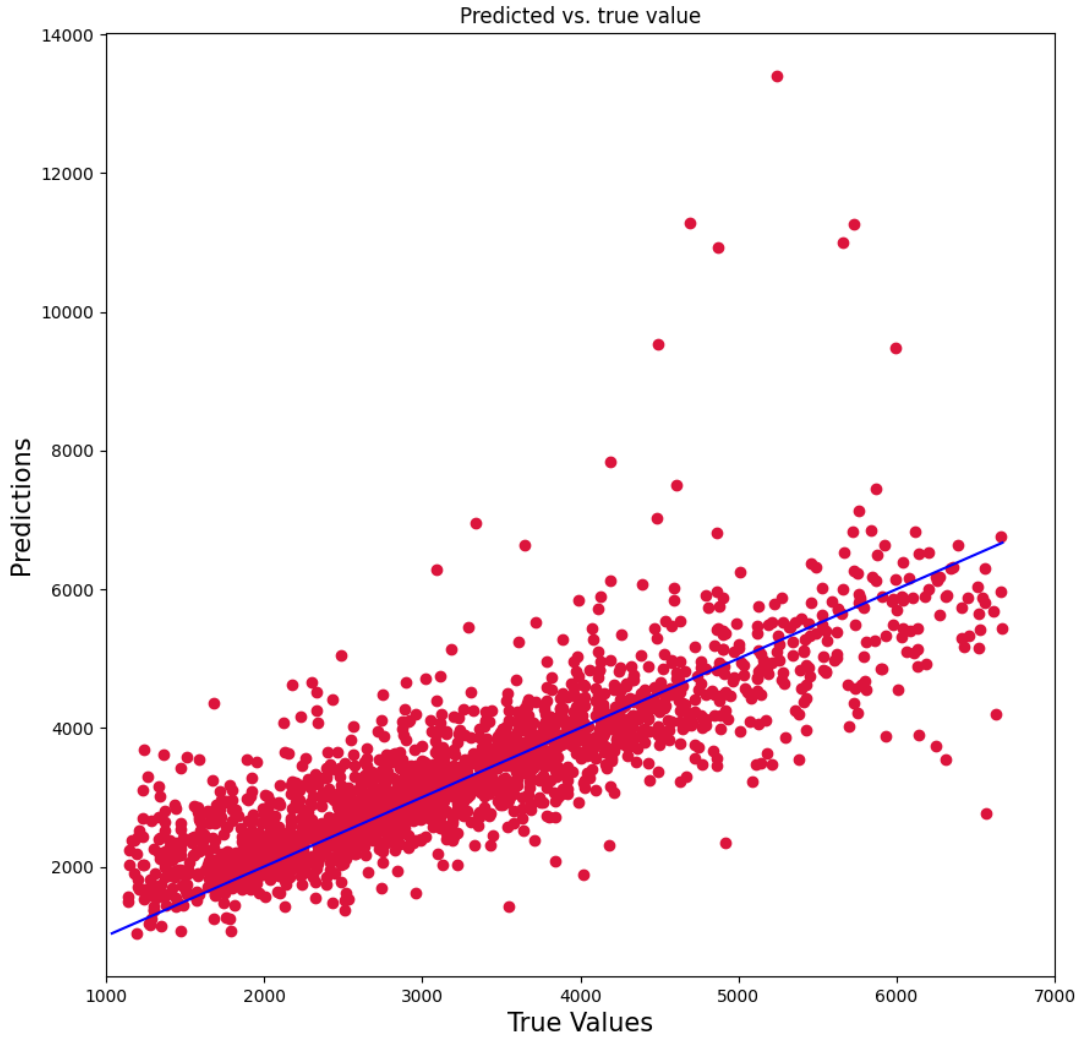


Figure 4.3: Regression Model Predictions

A relevant question for our proposed approach is the number of samples that need to be

manually digitized for the regression model to predict the target value accurately. Figure 4.4 shows the improvement in MAPE as the size of the training set increases. As the number of labeled samples in the training set increases from 3,000 to 8,000, the MAPE drops from roughly 18.6% to 17.5%. We did not collect additional samples, but based on the trend it appears that additional data would improve performance.

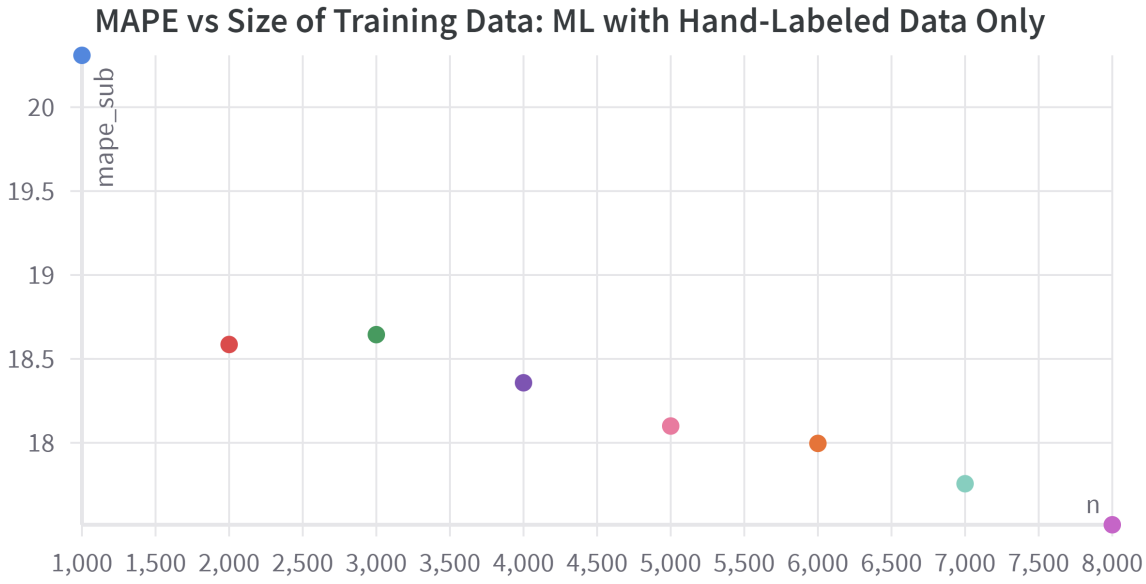


Figure 4.4: MAPE as size of hand-labeled training data increases

4.3 Augmented Regression Models

Next we explain the improvements in performance of the regression model by incorporating the predictions from the OCR model on unlabeled Hamilton County samples as training samples. As discussed in Section 4.2.2, in order to improve the accuracy of the OCR labels used for training the regression models, a threshold on the OCR model prediction confidence should be used. This presents a trade off between the quantity and quality of the training samples using this augmentation method. Choosing a high confidence threshold of the OCR predictions means the training samples are low but also contain fewer incorrect labels and vice versa. To examine

this effect, the performance of the augmented regression models using different OCR prediction confidence thresholds retaining the top 99%, 90% 75% and 50% of the most confident prediction, is shown in Figure 4.5.

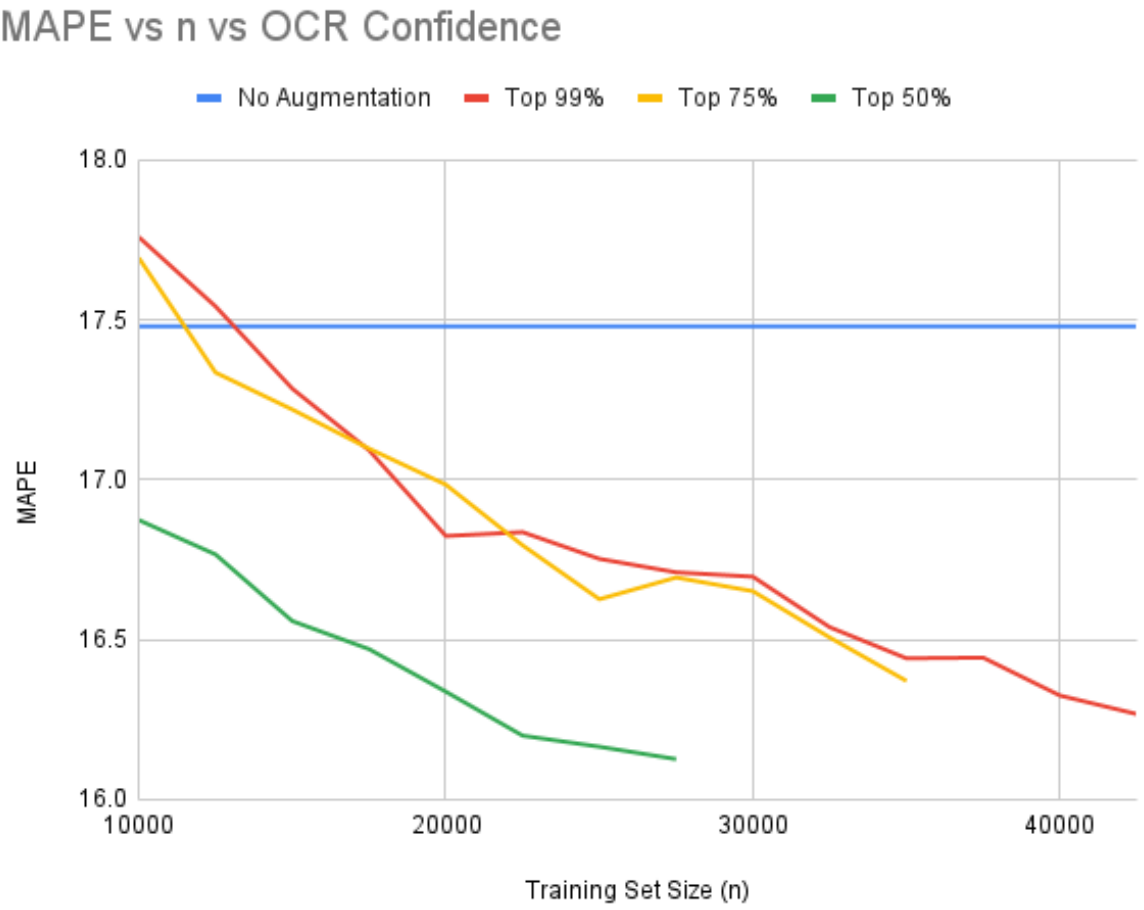


Figure 4.5: MAPE and OCR Confidence Threshold vs n

Compared to the regression model performance listed in Table 4.1 while we see a slight improvement in some accuracy measure such as MAPE from 17.48% to 16.12% and RMSPE from 27.73% to 24.01%, other measures MPE see a slight decline. We also note that while the amount of predictions within 20% of the true values improved from 73.44% to 74.55%, the amount of predictions within 5% and 10% of the true values decreased. This result suggest that

it is not conclusive that augmenting the regression models using OCR predictions is beneficial. From our analysis, the outliers in the OCR predictions, despite our efforts to remove them by applying a threshold for the OCR prediction confidence, are highly detrimental to regression models and offset the benefits of additional training samples. The predictions made by this mode is shown in Figure 4.6.

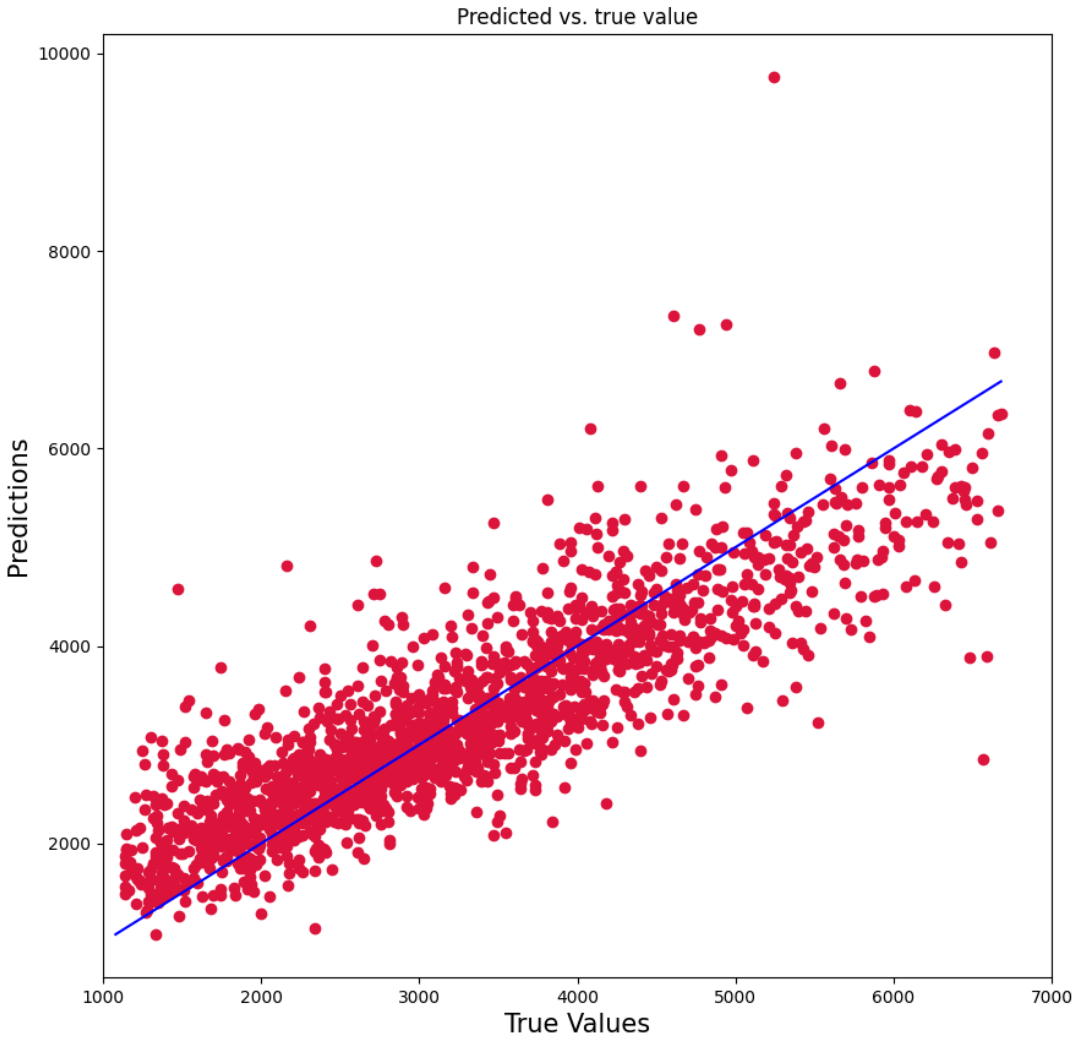


Figure 4.6: Augmented Regression Model Predictions

4.4 Generalization

We verify the generalization power of the regression model by evaluating its performance on three groups of data unseen during training. First, we evaluate the model's performance on Franklin County (Columbus) as a way to measure its prediction power on other counties in the US. Next, we also evaluate the model's performance on parcels for whose ownership cards we were unable to perform OCR to test our model's generalization on other parcels in the same county. Finally, we check whether our model is biased for parcels where we could not retrieve the ownership cards.

4.4.1 Franklin County

We observe that the distributions of the target values of the two counties are different with the median target value being $2,300 in Franklin County, which is lower than the Hamilton County median of $3,085. To correct for the difference between these two distributions we randomly sample 100 parcels in Franklin County and compute the mean and standard deviation of the two counties and apply the adjustment using Equation 4.1.

$$\begin{aligned} & Y_{Franklin} \\ & = \frac{Y_{Hamilton} - \mu_{Hamilton}}{\sigma_{Hamilton}} * \sigma_{Franklin} + \mu_{Franklin} \end{aligned} \tag{4.1}$$

The results for Franklin County are worse than those for Hamilton County test set, with an MAPE of 22.72%, which is a significant degradation, but still outperforms the baseline model. Figure 4.7 shows that the model predictions correlate poorly with the true values which suggests the model isn't capturing the underlying relations between the input features and the target value accurately.

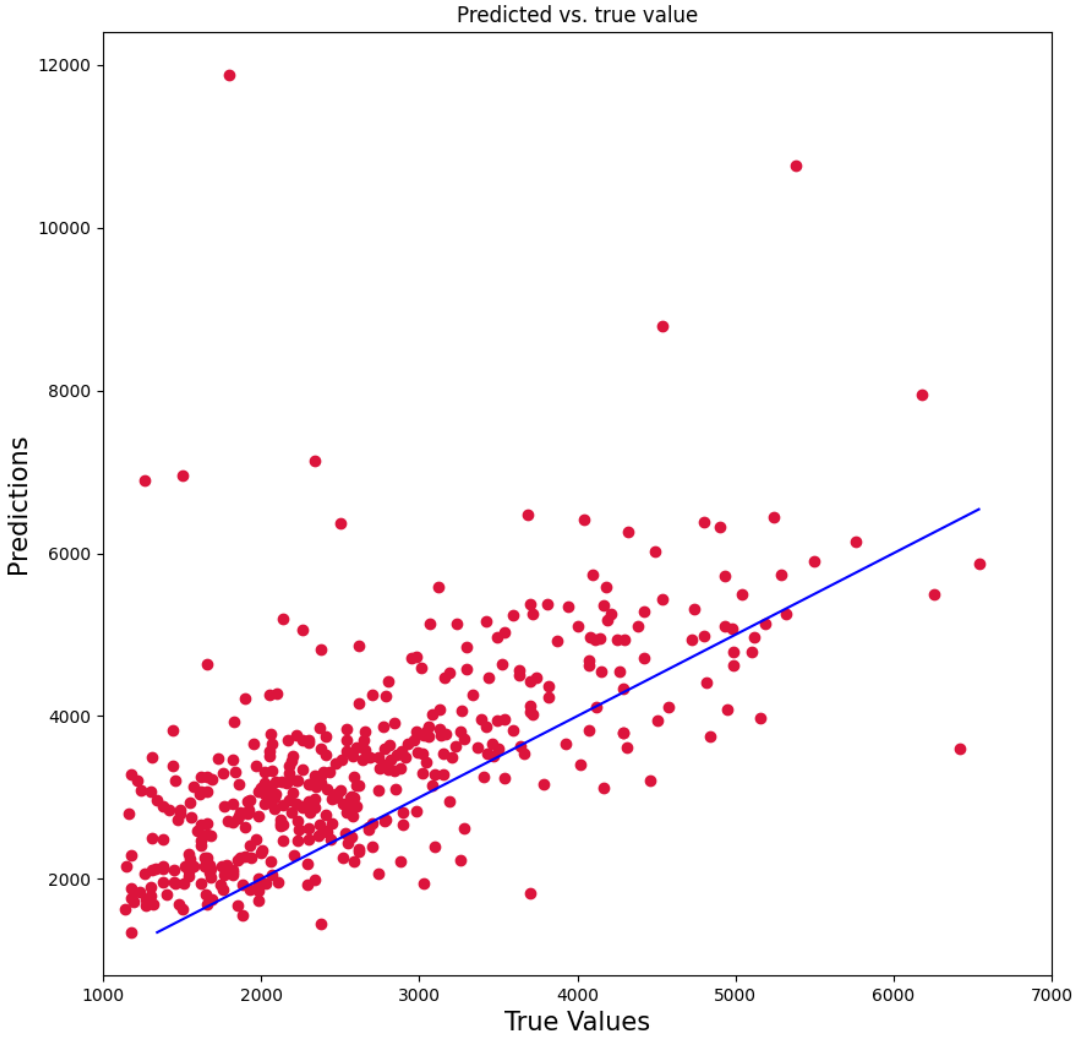


Figure 4.7: Regression Model Predictions on Franklin County

4.4.2 Hamilton County where OCR Methods Failed

The prediction statistics of using the augmented regression model on OCR segmentation failures is shown in Table 4.3. We observe no significant difference in prediction performance which confirms that we generalize well to this dataset.

Table 4.3: Regression Model Generalization on OCR Failures

Metrics	Augmented	OCR Failures
R^2 (higher is better)	0.7428	0.6671
MAPE (lower is better)	16.12%	15.98%
RMSPE (lower is better)	24.01%	25.18%
MPE (lower is better)	11.27%	11.56%
Within 5% of True Value (higher is better)	24.39%	23.68%
Within 10% of True Value (higher is better)	45.85%	45.82%
Within 20% of True Value (higher is better)	74.55%	73.79%

Chapter 5

Discussion

5.1 Cost Accuracy Trade-off

One of the main benefits of the proposed OCR and regression techniques for extracting values from historical records is the ability to scale to large numbers of documents with minimal cost. As a baseline we consider a hypothetical collection of 353,973 each with a single value to extract, which matches the number of properties as in Hamilton County. The cost and accuracy comparison of the two proposed methods is shown in Figure 5.1 For details on the following estimates calculations, see Appendix J.

To estimate the cost savings of the OCR methods, we assume scanned documents are available and compare the estimated costs of manual data entry of a single target value against adapting the OCR methods. Manually extracting a single value from scanned documents at the rate we used for the manual labeling process on 353,973 documents will cost an estimated $24,789.22. In contrast, the cost of employing a data scientist to adapt the OCR methods described in this work to a different document will cost $5,568.10, which is 22% of the manual process. The drawback for this cost reduction is the reduction in accuracy with an MAPE of 14.72%. Based on our experience with hand-labeling, given the structured nature of these documents, we assume that manual collection would yield close to perfect accuracy if clear instructions are provided and the

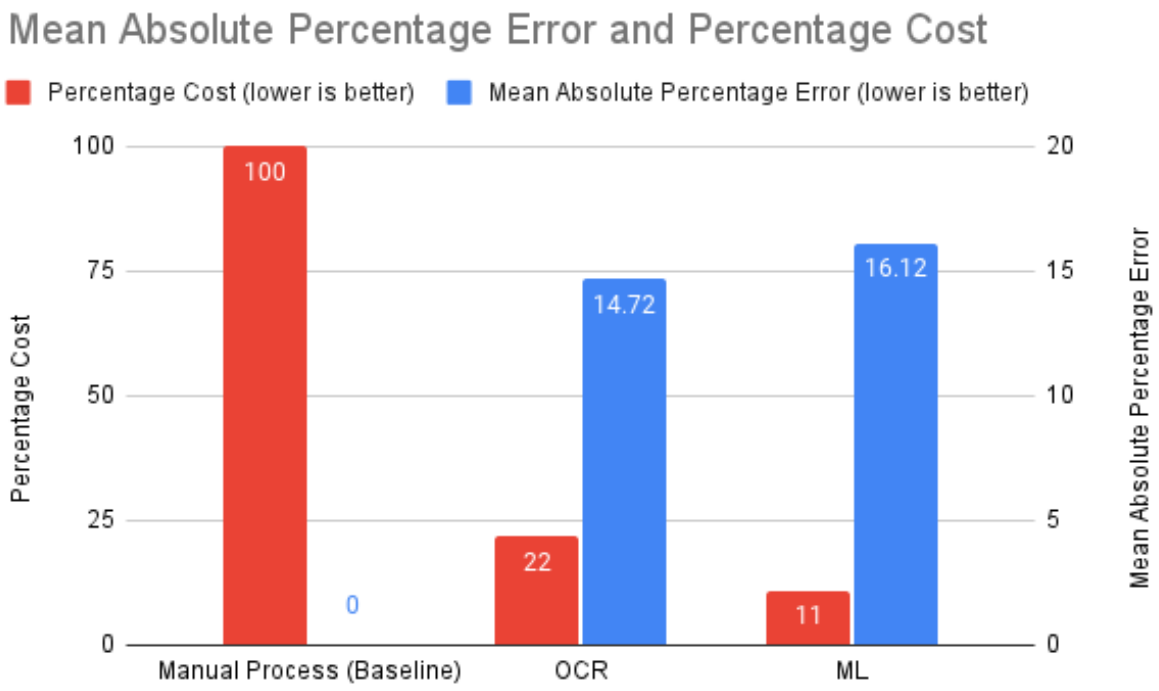


Figure 5.1: Cost and Accuracy Comparisons of Proposed Methods

work is well distributed. This assumption may not hold if the quality control of manual collection is difficult, and thus reduces the relative accuracy cost of OCR.

In the scenario where scanned documents are not available, we compared the cost of fully manual scanning and data entry process against the the proposed regression methods. Using online estimates of document scanning services, this will incur an average cost of $35,570.42 for 353,973 documents. Combined with the data entry costs listed previously, this gives a total cost of $60,359.64 for the manual process. We estimate that the cost to develop the regression model described in this work, including the costs of generating 12,423 training samples, to be $6,816.49. This represents a 11% of the cost of a comparable manual process but using this method further reduces the accuracy to an MAPE of 17.48%. There is a cost-accuracy tradeoff even within the regression method: as shown in Figure 4.4, one could incur a higher or lower cost of hand-labeling training samples based on the desired accuracy.

5.2 Future Work

Here we present several additional paths to explore for improving the results we presented here and address open questions.

5.2.1 OCR of Entire Historical Document

Our OCR methods currently only attempt to extract the initial building value as a proxy of the construction cost. This approach is ignoring a large amount of data present on the ownership documents including the value of the land or other components, changes in the valuation across the years and comments or details about the valuation changes. These data can be extracted by adapting our current segmentation technique to the entire document or using a more sophisticated deep learning based segmentation model to automatically recognize the positions and relationships in the tabular document. With additional time and computational resources, we can evaluate the feasibility of this approach.

5.2.2 Additional Contemporary Data and Training Samples for Regression Models

One way to improve our regression models is to improve the richness and quantity of the input data. From out analysis, we suspect that our regression models were limited due to missing input signals and the number of training samples we were able to generate using a manual process. For example, in our case study of estimating home values, additional contemporary data include home improvement permits and contemporary photos of the building. These data would help in bridging the gap between the contemporary information on the building and the changes in building value due to renovations and modifications as well as visual indications of how well the buildings have been maintained through the years. Extending this concept beyond the case study, we expect the performance of regression models to be sensitive to the quantity and quality

of input samples. Thus, gathering additional training samples and exploring additional sources of input features would be an important next step.

5.2.3 Deep Learning Models

The regression models we experimented with in this work are relatively simple models and do not take advantage of the latest development in deep learning. We chose to use these simpler models due to the simplicity of our input features and the small number of training samples. With a richer set of input features using additional sources of contemporary data along with more samples collected beyond Hamilton and Franklin counties, a more complex deep learning model may prove more successful at predicting the historical value of interest.

5.2.4 Domain adaptation

One of the challenge of using the methods proposed in this work to new data from another county is the cost of developing and training a new model for each new county. Ideally, we want to train one model and use it for all counties without degradation in prediction accuracy. However, we find through our evaluation of our regression model trained on Hamilton County on test data from Franklin county that generalization performance is poor. Assuming it is feasible to annotate a small set of training samples from the new target county, a more sophisticated few-shot domain adaptation approach could be considered. FOr example, a small number of samples from Franklin can be incorporated into the training data to improve generalization performance.

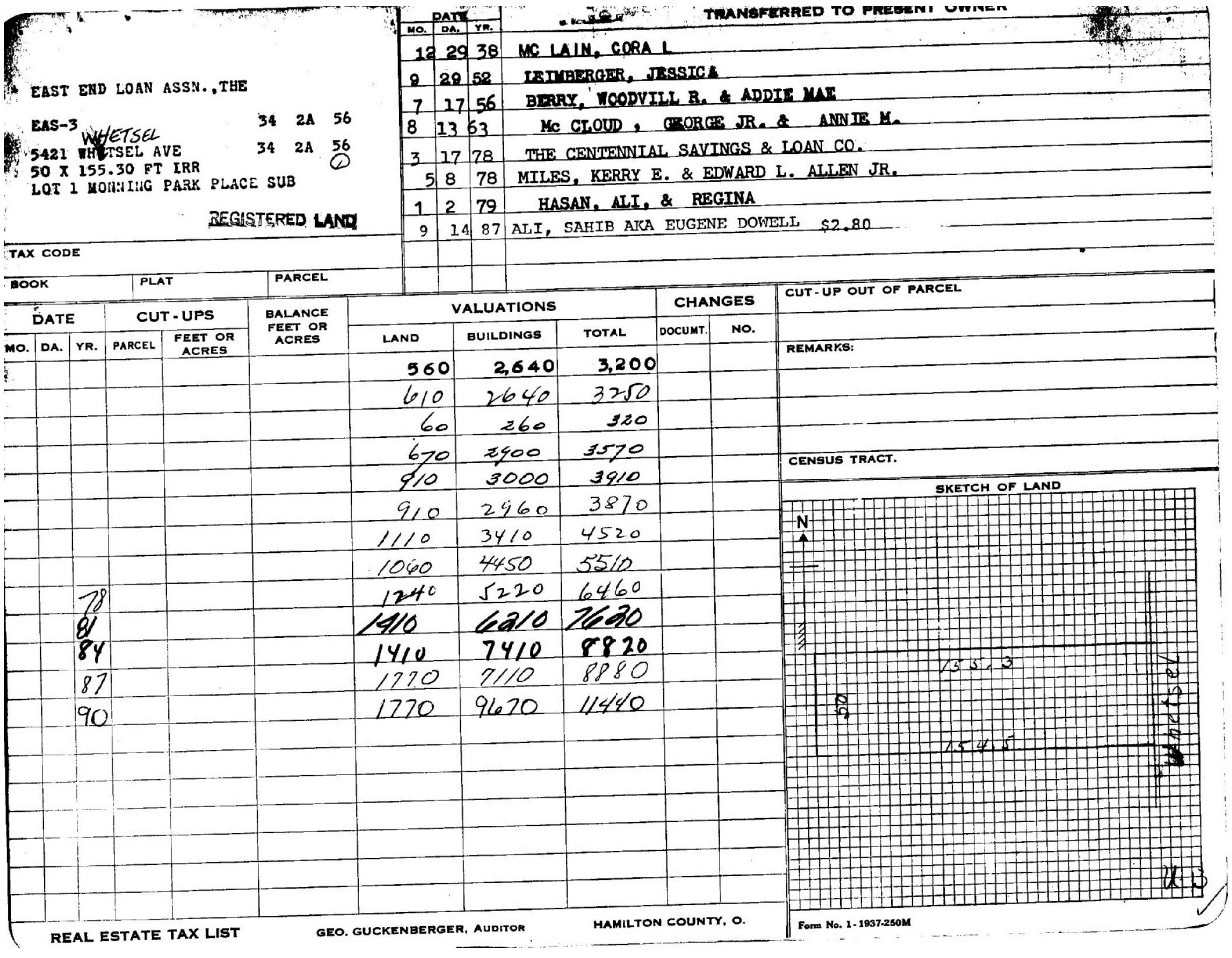
Chapter 6

Conclusion

Through this work we are able to show that machine learning and computer vision methods are viable approaches for digitizing data from tabular historical documents. For our chosen case study, these approaches a prediction accuracy of 14.72% MAPE and 17.48% MAPE, respectively. We also demonstrate that these methods are cost effective compared to existing manual methods, saving up to 78% with the OCR methods and 89% with regression methods. Though we show the feasibility of augmenting regression model training samples with OCR generated labels, additional work needs to be done to conclusively demonstrate its effectiveness. With potential improvements from expanding the complexity of the regression model, increasing the richness of the regression model inputs and applying our OCR methods on the full historical document, we expect our proposed methods to perform even better, given sufficient time and resources to explore these approaches. We hope our work highlights the benefits of using machine learning and computer vision in unlocking the wealth of data available in historical documents and serves as a guide for how these tools can be practically applied.

Appendix A

Sample Hamilton County Ownership Card



EAST END LOAN ASSN.,THE

EAS-3 WHETSEL 34 2A 56

5421 WHETSEL AVE 34 2A 56

50 X 155.30 FT IRR

LOT 1 MORNING PARK PLACE SUB

REGISTERED LAND

TAX CODE

BOOK | PLAT | PARCEL

DATE MO.	DA.	YR.	TRANSFERRED TO PRESENT OWNER
12	29	38	MC LAIN, CORA L
9	29	52	LEIMBERGER, JESSICA
7	17	56	BERRY, WOODVILL R. & ADDIE MAE
8	13	63	Mc CLOUD , GEORGE JR. & ANNIE M.
3	17	78	THE CENTENNIAL SAVINGS & LOAN CO.
5	8	78	MILES, KERRY E. & EDWARD L. ALLEN JR.
1	2	79	HASAN, ALI, & REGINA
9	14	87	ALI, SAHIB AKA EUGENE DOWELL $2.80

DATE MO.	DA.	YR.	CUT-UPS PARCEL	FEET OR ACRES	BALANCE FEET OR ACRES	VALUATIONS LAND	BUILDINGS	TOTAL	CHANGES DOCUMT.	NO.
						560	2,640	3,200		
						610	2640	3250		
						60	260	320		
						670	2900	3570		
						910	3000	3910		
						910	2960	3870		
						1110	3410	4520		
						1060	4450	5510		
		78				1240	5220	6460		
		81				1410	6210	7620		
		84				1410	7410	8820		
		87				1770	7110	8880		
		90				1770	9670	11440		

CUT-UP OUT OF PARCEL

REMARKS:

CENSUS TRACT.

SKETCH OF LAND

N

155.3

50

154.5

REAL ESTATE TAX LIST GEO. GUCKENBERGER, AUDITOR HAMILTON COUNTY, O.

Form No. 1-1937-250M

Figure A.1: Sample Ownership Card

Appendix B

Testing for Bias from Missing Ownership Cards

We wanted to confirm whether we introduced any bias in our regression models by ignoring the parcels which did not have any ownership cards available. Since we cannot evaluate the model's performance on ground truth values in these cases, we use a Classifier 2 Sample Test [18] using the contemporary features to check whether these cases are Missing At Random (MAR) to ensure we do not introduce any bias. We observe a p-value of 0.3870 from the test which confirms that these samples where ownership cards are missing are indeed MAR and does not introduce any bias in our models.

Appendix C

Manual Labeling

To ensure we have a reliable set of baseline labels for our models, we used Upwork to find contractor(s) to manually label a subset of our samples at a rate of up to 15 US$ an hour. We provided a total of 12,423 sample for which the first value in the ”BUILDING” column was recorded. Additional information such as the year this value was estimated as well as whether the value was handwritten were also recorded to distinguish whether the building values were the original value estimates from 1933. Finally we sample 1000 of the generated data and verify the correctness ourselves to ensure the accuracy of the labels before using it as ground truth for our models which showed that all manual labels we received were accurate.

Appendix D

Cleaning and processing of structured data from Hamilton County

Step 1: Load data

All raw data files were downloaded from source and placed into a Google Drive folder.

The data files were sourced from the Hamilton County Auditor's site downloads page, linked here. 'Tax Year Information Export' contains the tax assessment information, while both 'Historic Sales' and 'Building Information Export' contain building information.

Finally, we wrote a script to pull all the data from the Google Drive into a PostgreSQL database. All further processing happens in the database using SQL scripts.

Step 2: Fixing basic formatting issues

The first round of cleaning focused on fixing basic formatting and consistency issues. These include:

- Making the parcel identifier (parcelid) consistent across tables. For example, the parcelid had to be manually constructed in the older property transfer files by concatenating book, plat, parcel, and multi-owner (the fields that make up the parcelid) after removing special

characters. In other files, parcelids had to be converted to upper case.

- Standardizing NULL values. For example: in property class, null values were captured as two blankspace characters, while in property value the text 'New' was used.
- Optimizing the tables for query performance. We added indices on parcelid and converted string formats to numeric or datetime where possible.

We used this script to implement the cleaning, moving tables from a raw schema to a 'cleaned' schema in the database.

Step 3: Data quality issues and fixes

Table D.1: List of data quality issues and resolutions

Issue	**Decision**
Property class is captured in multiple tables, with inconsistent values for the same parcel	Use the tax assessment value, because it is the most updated source
Some parcels do not merge across tables. E.g., building info has 289 parcelids that don't merge to tax assessment	Drop rows in other tables that don't merge to tax assessment, as it is the most updated source.
Parcelids have duplicates because of multiple buildings on a parcel	For now, only analyse parcels with one building. Going forward, reshape data to wide format at parcel level, retaining info about multiple buildings.
Some buildings have 0 total square footage	For cases where other square footage fields are nonzero (e.g. floor 1, attic), impute value by summing these up. For buildings where all square footage columns are 0, drop rows because these buildings are torn down.

Once the basic cleaning was done we performed a more comprehensive data exploration. This raised further issues and inconsistencies which required discussion and decisions on how to handle such cases. These are summarized in Table D.1:

Step 4: Generating features

The following were the main types of transformations we did to the existing columns to create usable features:

- Group categorical variables with many closely related categories: e.g. combining Exceptional, Exceptional+, Outstanding and Extraordinary grades into 'Exceptional'.
- Creating categories from numeric features (e.g., categories of 'No attic', 'Partial attic', and 'Full attic' from attic square footage) and numeric features from categories (e.g., translating grade into a numeric scale). We did this for two reasons. First, we wanted to experiment with different feature representations to see how it would affect performance (rather than relying on the model to learn all patterns in the data). Second, some of these transformations were required to make the features standard across Hamilton and Franklin county.

We used this script to implement the additional cleaning and feature generation, moving tables from the 'cleaned' schema to 'processed'. The 'processed' schema is the final cleaned data fed as inputs to the modeling pipeline.

Appendix E

Standardizing features across Hamilton and Franklin County

Table E.1: Reconciling Contemporary Data for Hamilton and Franklin Counties

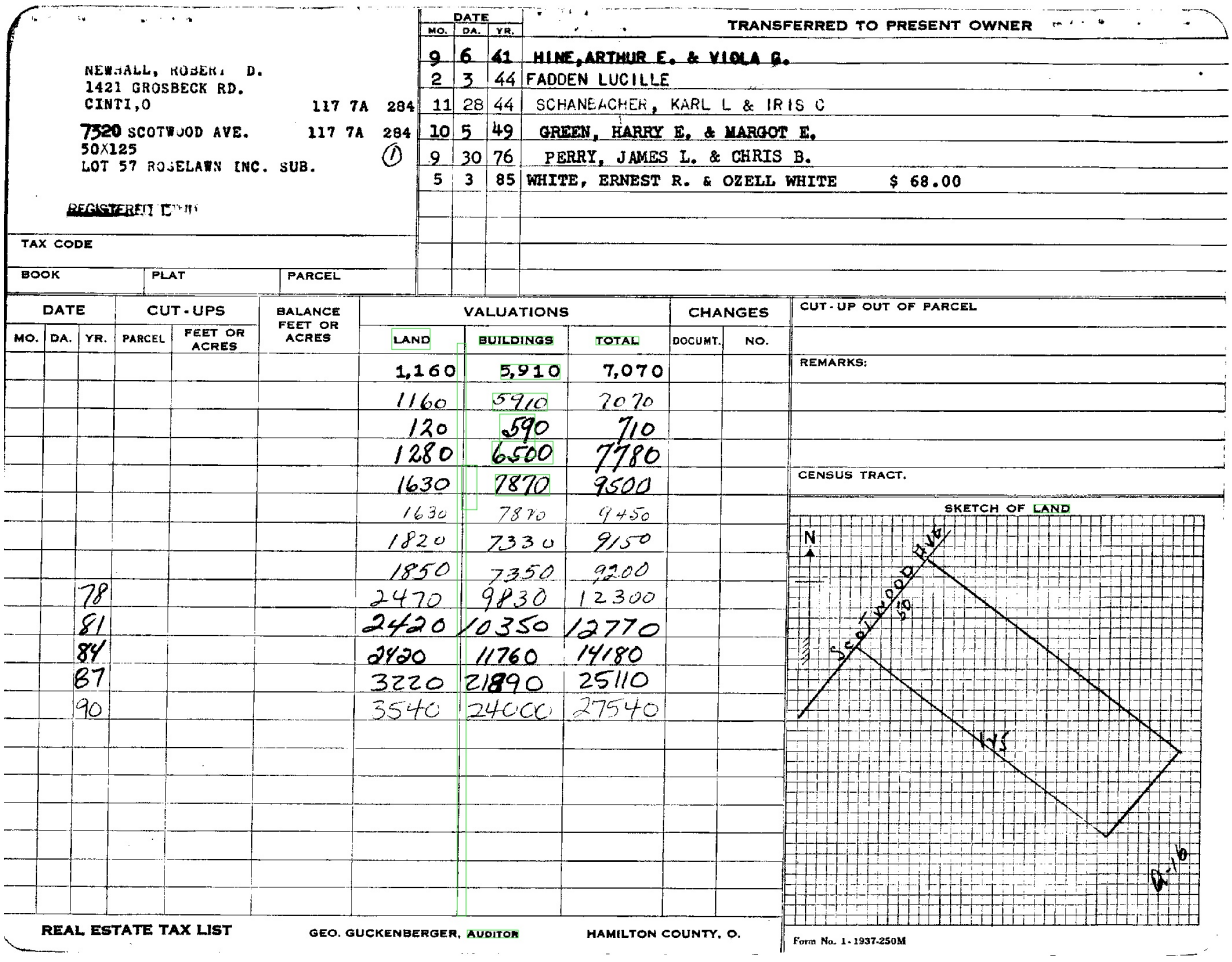
Issue	**Decision**
Information does not exist/is not captured at all by Franklin: e.g., half-floor and floor 2 square footage	Do not use these features in the generalized version of the model
Some information is captured at a higher or lower granularity. E.g., exact attic square footage is captured in Hamilton, but only broad categories are captured in Franklin (No Attic, Full Attic, Partial Attic).	Recode information to match the lowest granularity (e.g., convert attic square footage to categories based on logic)
Some information is captured in a different format or with different coding. E.g., grade descriptions are letter categories (e.g., A+2, AA-) rather than 'Outstanding'	Change Franklin coding to be consistent with Hamilton's

In order to test how well our regression model trained on Hamilton County generalizes to Franklin County, we needed to ensure that the features were standardized such that the model could be applied on the Franklin test set directly. Table E.1 notes the main types of differences between the two counties' contemporary data, and how we addressed it.

The final set of features used in the limited, generalizable model are: *attic category, living area square footage, floor 1 square footage, number of stories, year built, property use code, number of parcels per last sale, grade, exterior wall type, basement type, heating type, air conditioning type, total rooms, full bathrooms, half bathrooms, fireplaces, garage capacity*

Appendix F

Segmentation



NEWHALL, ROBERT D.
1421 GROSBECK RD.
CINTI,O 117 7A 284

7520 SCOTWOOD AVE. 117 7A 284
50X125
LOT 57 ROSELAWN INC. SUB. (1)

REGISTERED

TAX CODE

BOOK | PLAT | PARCEL

DATE MO.	DA.	YR.	TRANSFERRED TO PRESENT OWNER
9	6	41	HINE, ARTHUR E. & VIOLA G.
2	3	44	FADDEN LUCILLE
11	28	44	SCHANEACHER, KARL L & IRIS C
10	5	49	GREEN, HARRY E. & MARGOT E.
9	30	76	PERRY, JAMES L. & CHRIS B.
5	3	85	WHITE, ERNEST R. & OZELL WHITE $ 68.00

DATE MO.	DA.	YR.	CUT-UPS PARCEL	FEET OR ACRES	BALANCE FEET OR ACRES	VALUATIONS LAND	BUILDINGS	TOTAL	CHANGES DOCUMT.	NO.
						1,160	5,910	7,070		
						1160	5910	7070		
						120	590	710		
						1280	6500	7780		
						1630	7870	9500		
						1630	7870	9450		
						1820	7330	9150		
						1850	7350	9200		
	78					2470	9830	12300		
	81					2420	10350	12770		
	84					2420	11760	14180		
	87					3220	21890	25110		
	90					3540	24000	27540		

CUT-UP OUT OF PARCEL

REMARKS:

CENSUS TRACT.

SKETCH OF LAND

REAL ESTATE TAX LIST GEO. GUCKENBERGER, AUDITOR HAMILTON COUNTY, O.

Form No. 1-1937-250M

Figure F.1: Sample TesseractOCR Output

This task involves recognizing the column header “Buildings” in the image and extracting the bounding boxes of the first cell below it. In this work, we are concerned with extracting the initial construction cost of the building for which we deem the first entry under the ”Buildings” column to be a good proxy.

For the task of locating each cell segment, we begin with TesseractOCR as a baseline to label the bounding boxes for sequences of letters and digits. However, this proved to be difficult since there were many false positives and negatives.

Here we can see several issues. First, there are false positives where non digit elements such grid lines being recognized as characters by TesseractOCR. Second there are false negatives where digits further down the column are not recognized. Furthermore, some sequences of characters are not fully recognized. For example only the ”59” of the ”590” sequence is recognized. Finally the recognized characters are not always correct. For example, the first three rows were recognized as ”5,910”, ”SULO” and ”Ff” of which only the first row is correct. Given that TesseractOCR is a pretrained model, we found it difficult to modify its behavior for our particular problem and proceeded with building our own solution.

For the first step, we retain the use of TesseractOCR for locating the ”Buildings” column header and creating a cropped image around the column header. For example of the cropped document containing the detected column header, see Figure F.2.

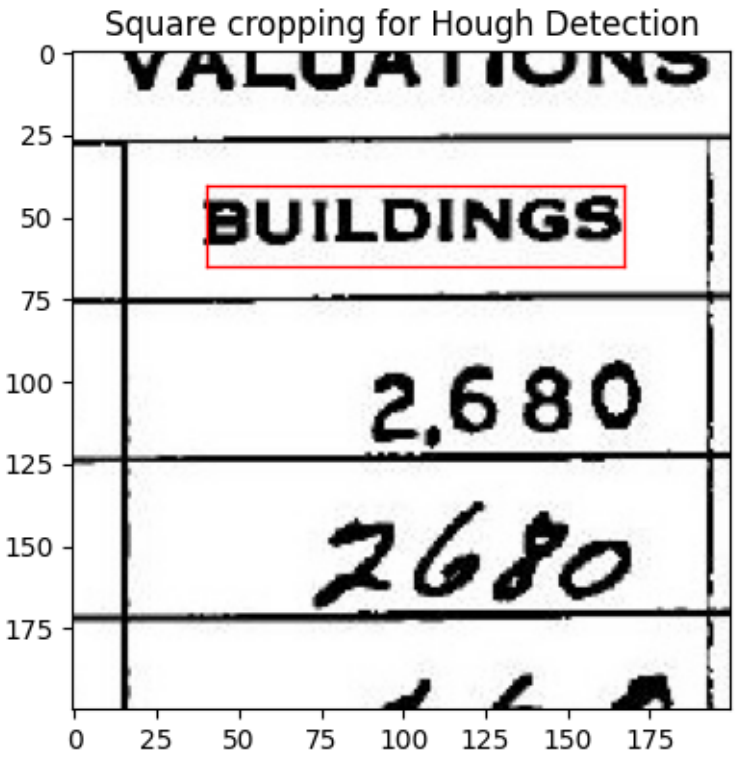


Figure F.2: Sample cropped document

To extract the cells below the header, we then use Hough Transform [8] to detect the main line segments in the cropped image. An example of the document with detected lines overlaid on top is shown in Figure F.3.

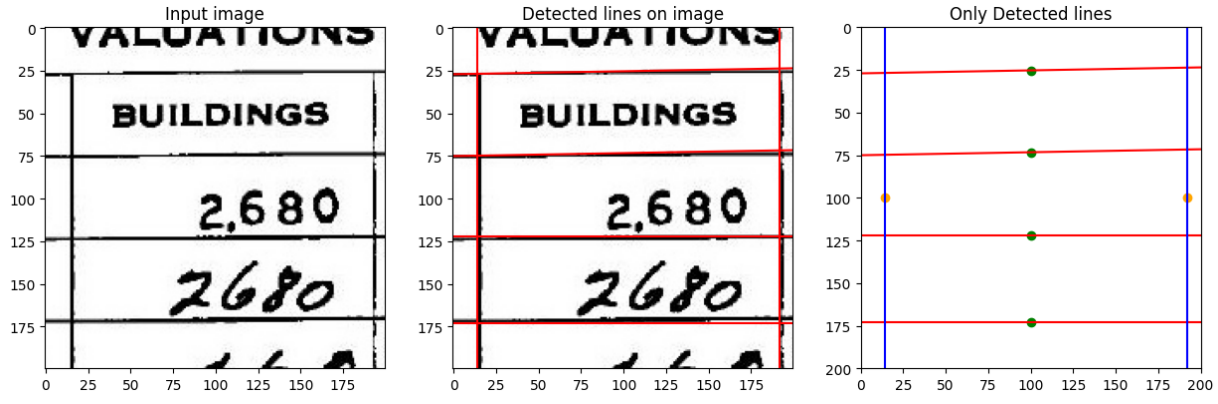


Figure F.3: Sample line detection using Hough Transform

Finally, we use the detected lines and compute the intersections to determine the corners containing the cell we are interested in, which is then used to create a final image of the cell stretched to be a regular rectangle, see Figure F.4.

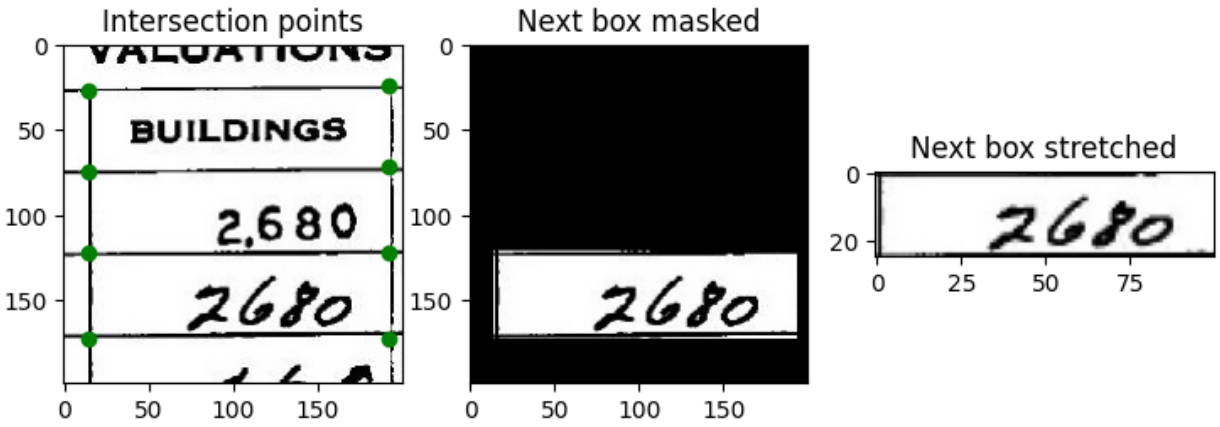


Figure F.4: Extracting a sample cell as a rectangular image

The final output is then ready to be used as an in put to OCR models.

Appendix G

OCR models

For the OCR task, we aim to retrieve a numeric value from the segments collected by the process described in the previous section. We experiment with both TesseractOCR and TrOCR to detect numbers and found the results of TrOCR to be significantly better than those obtained with TesseractOCR.

G.1 TesseractOCR

Our initial experiments with TesseractOCR involved using it for both segmentation and OCR since it outputs the bounding boxes, characters detected as well as its confidence of the predictions. This is promising since it provides all of the required information for constructing a structured output for tabular data. However, we quickly found that TesseractOCR is trained to be a general OCR tool that also recognizes letters and punctuation in addition to the digits that we are interested in and often confuses between them. Furthermore, TesseractOCR performs especially poorly on handwritten digits. As a result, we found that we needed to do significant amount of post-processing to retrieve any meaningful results. Even with all of the processing we were still only able to accurately retrieve the target value in 52.5% of our test cases, see Figure G.1 for the example predictions. Given these poor results we abandoned further work using this

tool for the OCR task.

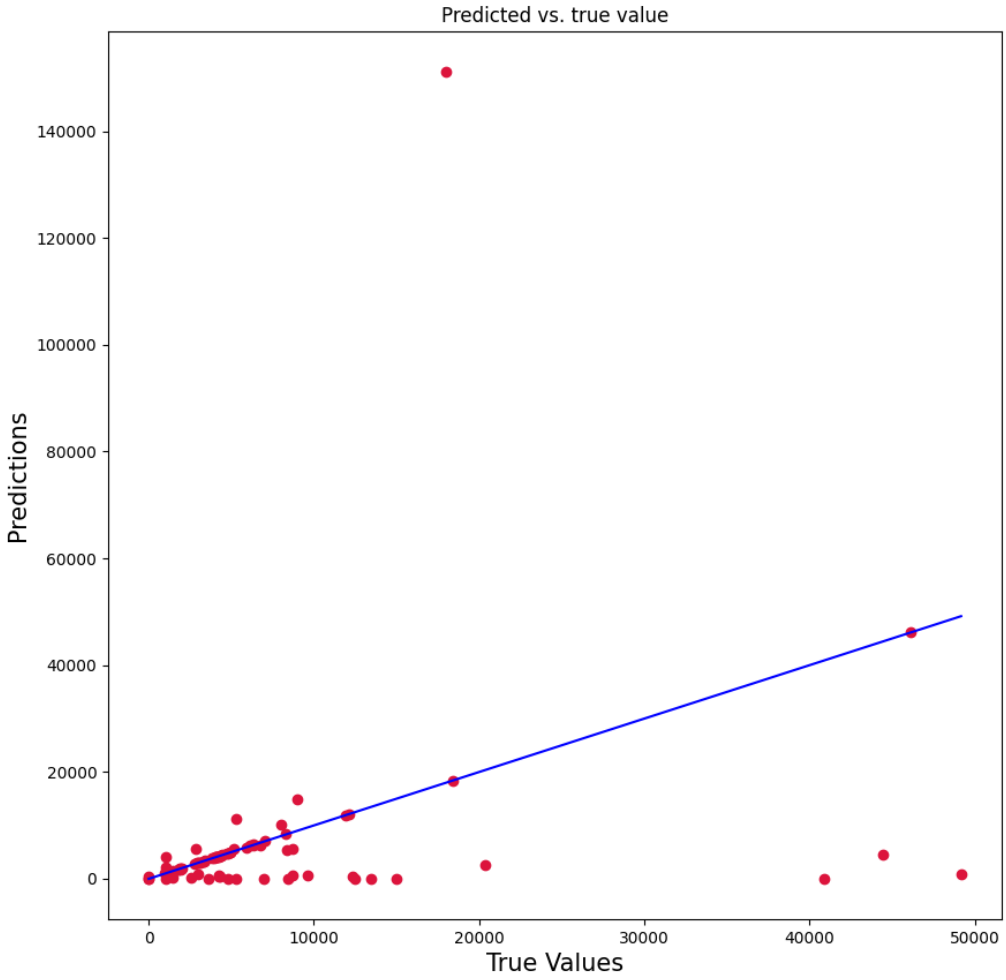


Figure G.1: TesseractOCR predictions

G.2 TrOCR

Our experiments with the TrOCR model is more successful. While the pre-trained TrOCR model suffers from similar errors as TesseractOCR such as recognizing letters and punctuation in addition to the digits we are interested in, we found that even with minimal fine-tuning on 500

training samples, we can achieve up to 95% exact match in our test set, a drastic improvement over TesseractOCR. Analysing the errors suggested that TrOCR was performing poorly on handwritten digits due to the lack of training samples containing handwriting. To address this deficiency, we combined our training set with the CAR-B dataset [7] of handwritten digit strings from checks to our training samples and surpassed the performance of TrOCR trained on only our dataset or only on CAR-B. A table of the performance of our TrOCR fine tuning experiments is found in Table G.1.

Table G.1: TrOCR Fine-tuning experiments

Fine-tuning Experiments	Exact match accuracy
Our Dataset n=500 (3 iters)	95%
CAR-B n=3k (3 iters)	4.90%
Our Dataset n=5k (3 iters)	97.17%
Our Dataset n=7k combined with CAR-B n=3k (3 iters)	**98.69%**
CAR-B n=3k (3 iters) then Our Dataset n=7k (3 iters)	95.51%

Further ablation studies on hyperparameters for TrOCR fine-tuning iterations did not yield significant improvements and we selected our best performing experiment as the model used to report our results.

Appendix H

Model class selection

Table H.1: Performance of regression model classes (no tuning)

Model Class	RMSE
Poisson Regressor	1068.42
Random Forest Regressor	1103.62
Huber Regressor	1117.24
Gamma Regressor	1144.69
XGB Regressor	1226.48
LassoLarsCV	1229.35
Gradient Boosting Regressor	1243.21
Lasso	1255.50
Light GBM Regressor	1271.28
ElasticNet	1303.20
Ridge	1432.39
Linear Regression	1444.08
Decision Tree Regressor	1681.67
AdaBoost Regressor	1681.67

Results of a preliminary search for promising model classes to conduct hyperparameter searches on.

Appendix I

Feature Importance

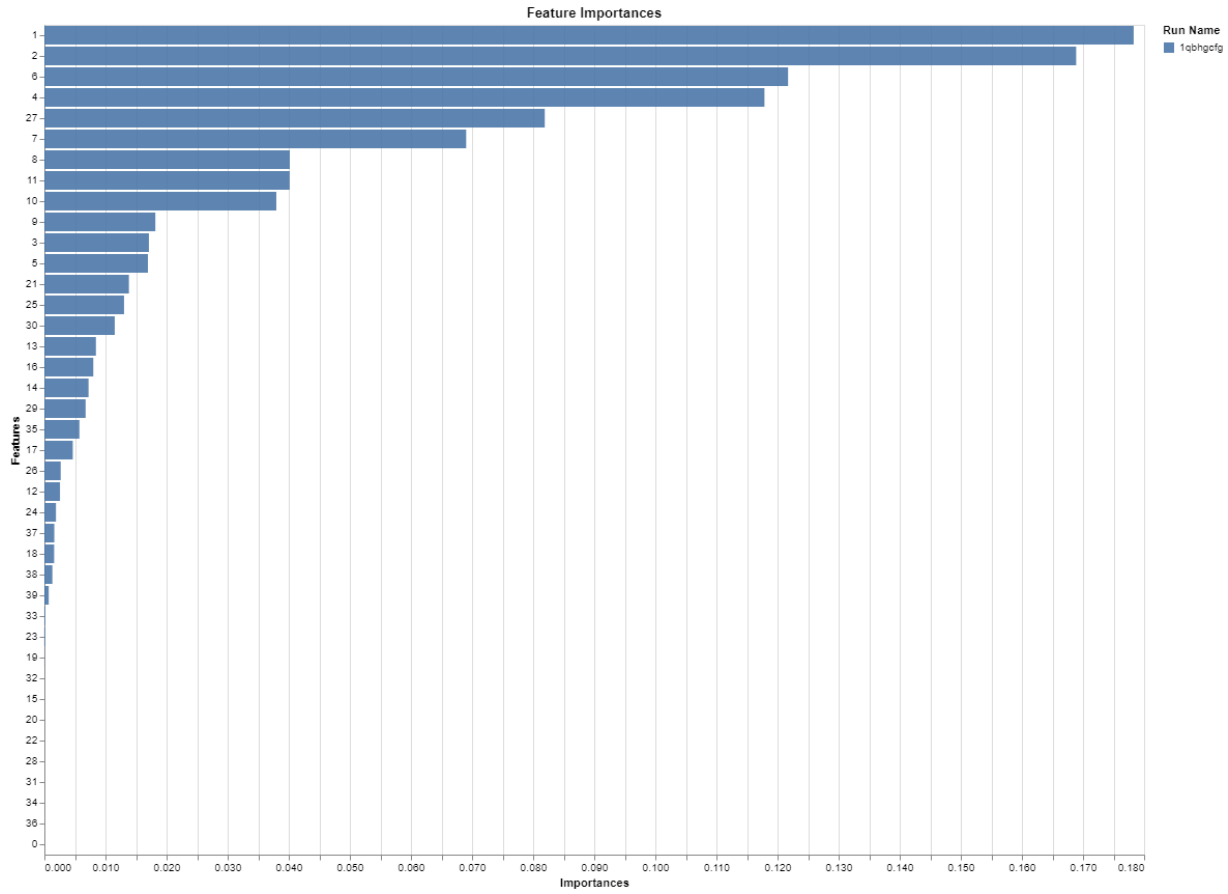


Figure I.1: Feature Importances: ML Model (without OCR augmentation)

Appendix J

Cost estimation

We find that most digital record services that offer data entry of specific values in the document to involve two steps, like at Iron Mountain [20]. Typically, the document is first scanned, then OCR or manual entry is performed on the scanned document. This workflow is also used in previous research into historical document digitization [29]. As such we estimate cost of the two steps individually as part of calculations.

For the estimation of scanning 353,973 pages of ownership documents we use the online estimators from two separate services. SecureScan [25] gives a quote of $45,477.80 and ILM Corp [5] gives a quote of $25,663.04, giving an average estimated cost of $35,570.42 or $0.10049 per document.

For the estimation of hiring contractors to extract the initial construction costs from scanned documents we use the same rate as our manual labeling contract on Upwork. In our case, we charged a rate of $15/hr and was able to label 12,423 samples in 58 hours. Extrapolating from this rate to 353,973 gives an estimated cost of $24,789.22 or $0.07003 per document.

We then estimate the cost of developing the OCR and regression models. Considering the time to develop the two proposed models were comparable and required one 14-week semester of work at an estimated 12 hours per week, it took about 84 hours to develop each individual model. Using an estimate of an average Data Scientist salary of $55.93 from Indeed.com [15],

we estimate the cost of developing each model at $4698.12.

For both methods, additional costs need to be included for generating the training labels. For the OCR methods which correspond to the scenario where documents are scanned, only the data entry costs are involved which sums to $869.98 for 12,423 training samples. This gives a final cost for OCR methods of $5568.10.

For our regression model, we need to collected 12,423 training samples from documents that are not scanned. Using the scanning and data entry costs per document listed above this would add an additional $2118.37 to the development of the regression model giving a total of $6816.49.

Bibliography

[1] Hamilton County Auditor. Hamilton county auditor: Real estate tax valuation. https://www.hamiltoncountyauditor.org/revalue.asp, 2023. Accessed: 2023-04-23. 2.1

[2] Martha J. Bailey, Susan H. Leonard, Joseph Price, Evan Roberts, Logan Spector, and Mengying Zhang. Breathing new life into death certificates: Extracting handwritten cause of death in the life-m project. *Explorations in Economic History*, 87:101474, 2023. ISSN 0014-4983. doi: https://doi.org/10.1016/j.eeh.2022.101474. URL https://www.sciencedirect.com/science/article/pii/S0014498322000523. Methodological Advances in the Extraction and Analysis of Historical Data. 1

[3] Alejandro Baldominos, Iván Blanco, Antonio José Moreno, Rubén Iturrarte, Óscar Bernárdez, and Carlos Afonso. Identifying real estate opportunities using machine learning. *Applied sciences*, 8(11):2321, 2018. 1.1

[4] Callum Booth, Robert Shoemaker, and Robert Gaizauskas. A language modelling approach to quality assessment of OCR'ed historical text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5859–5864, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.630. 1.1

[5] ILM Corp. ILM Corp cost of document scanning. https://www.ilmcorp.com/

tools-and-resources/cost-of-document-scanning/, 2023. Accessed: 2023-04-10. J

[6] Sergio Correia and Stephan Luck. Digitizing historical balance sheet data: A practitioner's guide. *Explorations in Economic History*, 87:101475, 2023. ISSN 0014-4983. doi: https://doi.org/10.1016/j.eeh.2022.101475. URL https://www.sciencedirect.com/science/article/pii/S0014498322000535. Methodological Advances in the Extraction and Analysis of Historical Data. 1.1

[7] Markus Diem, Stefan Fiel, Florian Kleber, Robert Sablatnig, Jose M. Saavedra, David Contreras, Juan Manuel Barrios, and Luiz S. Oliveira. Icfhr 2014 competition on handwritten digit string recognition in challenging datasets (hdsrc 2014). In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 779–784, 2014. doi: 10.1109/ICFHR.2014.136. 3.1.2, G.2

[8] Richard O. Duda and Peter E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM*, 15(1):11–15, jan 1972. ISSN 0001-0782. doi: 10.1145/361237.361242. URL https://doi.org/10.1145/361237.361242. F

[9] James J. Feigenbaum. Automated census record linking: A machine learning approach. Accessed: 2023-04-23, 2016. 1.1

[10] Pascal Fischer, Alen Smajic, Giuseppe Abrami, and Alexander Mehler. Multi-type-td-tsr–extracting tables from document images using a multi-stage pipeline for table detection and table structure recognition: From ocr to structured table representations. In *KI 2021: Advances in Artificial Intelligence: 44th German Conference on AI, Virtual Event, September 27–October 1, 2021, Proceedings 44*, pages 95–108. Springer, 2021. 1.1

[11] Michel Fouquin and Jules Hugot. Two centuries of bilateral trade and gravity data: 1827-2014. Accessed: 2023-04-23, 2016. 1

[12] Winky KO Ho, Bo-Sin Tang, and Siu Wai Wong. Predicting property prices with machine

learning algorithms. *Journal of Property Research*, 38(1):48–70, 2021. 1.1

[13] Junia Howell. Reimagining equity: Towards an equitable economic model for social housing. Vienna International Summer School on Social Housing Production, 2022. URL https://iba-researchlab.at/summer-school-2022/. 1

[14] Junia Howell and Elizabeth Korver-Glenn. The Increasing Effect of Neighborhood Racial Composition on Housing Values, 1980–2015. *Social Problems*, 68(4):1051–1071, 09 2020. ISSN 0037-7791. doi: 10.1093/socpro/spaa033. URL https://doi.org/10.1093/socpro/spaa033. 1

[15] Indeed.com. Indeed data scientist salary in united states. https://www.indeed.com/career/data-scientist/salaries, 2023. Accessed: 2023-04-13. J

[16] Huseyin Kusetogullari, Amir Yavariabdi, Johan Hall, and Niklas Lavesson. Digitnet: A deep handwritten digit detection and recognition method using a new historical handwritten digit dataset. *Big Data Research*, 2020. 3.1.2

[17] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models, 2021. URL https://arxiv.org/abs/2109.10282. 1.1

[18] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests, 2018. B

[19] Jiří Martínek, Ladislav Lenc, and Pavel Král. Training strategies for ocr systems for historical documents. In John MacIntyre, Ilias Maglogiannis, Lazaros Iliadis, and Elias Pimenidis, editors, *Artificial Intelligence Applications and Innovations*, pages 362–373, Cham, 2019. Springer International Publishing. ISBN 978-3-030-19823-7. 1.1

[20] Iron Mountain. Iron Mountain document scanning & digital storage services. https://www.ironmountain.com/services/document-scanning-and-digital-storage#howitworks, 2023. Accessed: 2023-04-13. J

[21] Jonas Mueller-Gastell, Marcelo Sena, and Chiin-Zhe Tan. A multi-digit ocr system for historical records (computer vision). Accessed: 2023-04-13, 2020. 1.1

[22] Smita Pallavi, Raj Ratn Pranesh, and Sumit Kumar. A conglomerate of multiple OCR table detection and extraction. *CoRR*, abs/2010.08591, 2020. URL https://arxiv.org/abs/2010.08591. 1.1

[23] Joseph Price, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley. Combining family history and machine learning to link historical records. Technical report, National Bureau of Economic Research, 2019. 1.1

[24] Ashish Ranjan, Varun Nagesh Jolly Behera, and Motahar Reza. *OCR Using Computer Vision and Machine Learning*, pages 83–105. Springer International Publishing, Cham, 2021. ISBN 978-3-030-50641-4. doi: 10.1007/978-3-030-50641-4_6. URL https://doi.org/10.1007/978-3-030-50641-4_6. 1.1

[25] Secure Scan. Secure Scan document scanning price calculator. https://www.securescan.com/document-scanning-price-calculator/, 2023. Accessed: 2023-04-10. J

[26] R. Smith. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633, 2007. doi: 10.1109/ICDAR.2007.4376991. 1.1

[27] Dieudonné Tchuente and Serge Nyawa. Real estate price estimation in french cities using geocoding and machine learning. *Annals of Operations Research*, pages 1–38, 2022. 1.1

[28] Bogdan Trawiński, Zbigniew Telec, Jacek Krasnoborski, Mateusz Piwowarczyk, Michał Talaga, Tedeusz Lasota, and Edward Sawiłow. Comparison of expert algorithms with machine learning models for real estate appraisal. In *2017 IEEE international conference on innovations in intelligent systems and applications (INISTA)*, pages 51–54. IEEE, 2017. 1.1

[29] Amir Yavariabdi, Huseyin Kusetogullari, Turgay Celik, Shivani Thummanapally, Sakib

Rijwan, and Johan Hall. Cardis: A swedish historical handwritten character and word dataset. *IEEE Access*, 10:55338–55349, 2022. doi: 10.1109/ACCESS.2022.3175197. 1, 1.1, J

[30] Yun Zhao, Girija Chetty, and Dat Tran. Deep learning with xgboost for real estate appraisal. In *2019 IEEE symposium series on computational intelligence (SSCI)*, pages 1396–1401. IEEE, 2019. 1.1