

Automating Workflows for the City of Montgomery

BRANNON WALDEN



Contents

1. Introduction
2. Explanation of Data
3. Feature Engineering
4. Model Analysis
5. Model Summary
6. Next Steps
7. Appendix



Introduction

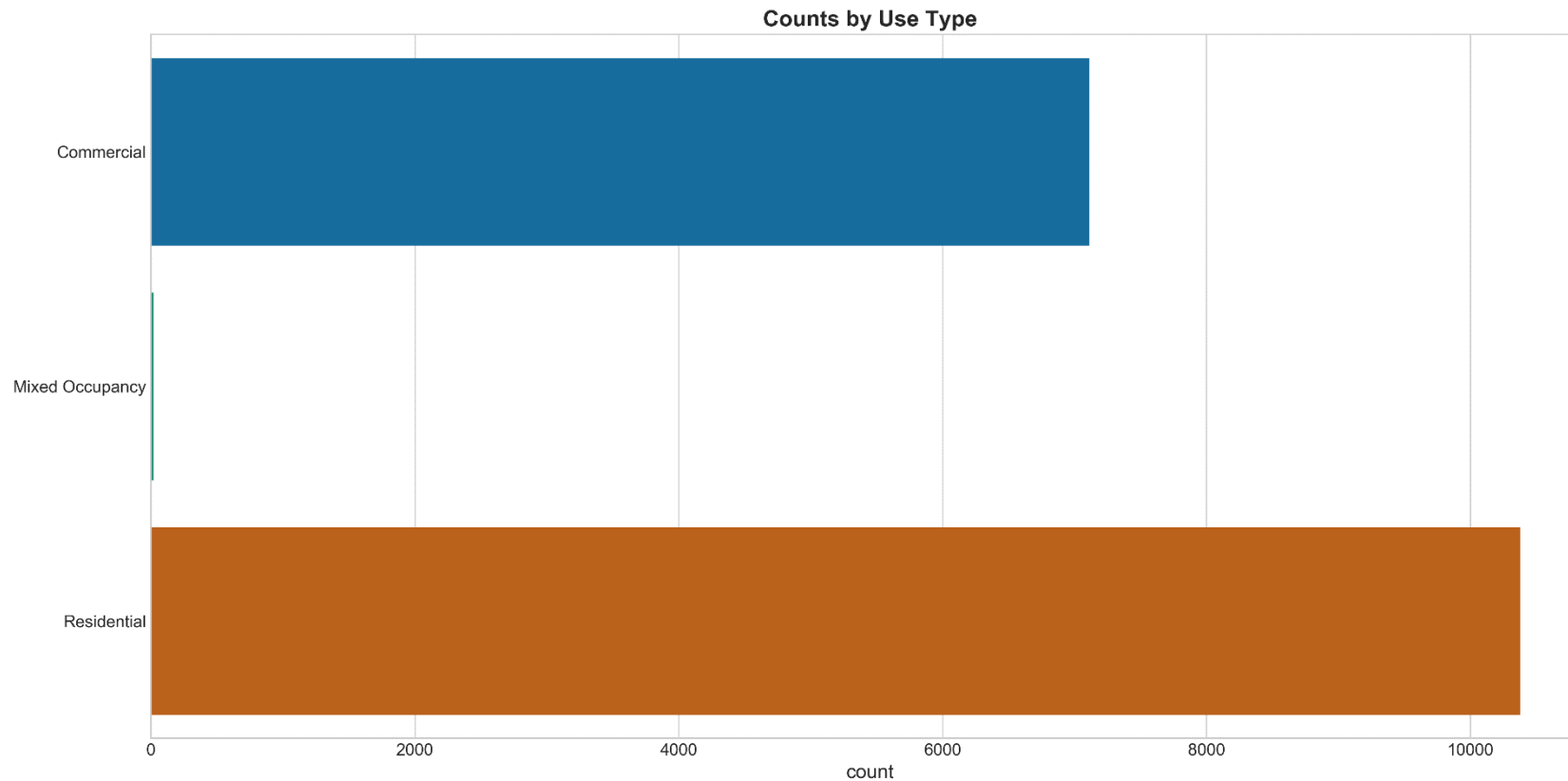
The City of Montgomery Inspections Department:

- Issues building permits for construction projects.
- Documents permit applications by entering long descriptions of the projects scope along with many other fields such as location, owner, estimated costs, permit types, zoning, etc.
- Data entry is time consuming and diverts staff attention away from code enforcement and other tasks.

Hypothesis:

The Use Type field, and possibly Job Type, could be automagically populated without human input and with better than average (>52.81%) results using natural language processing and machine learning techniques.

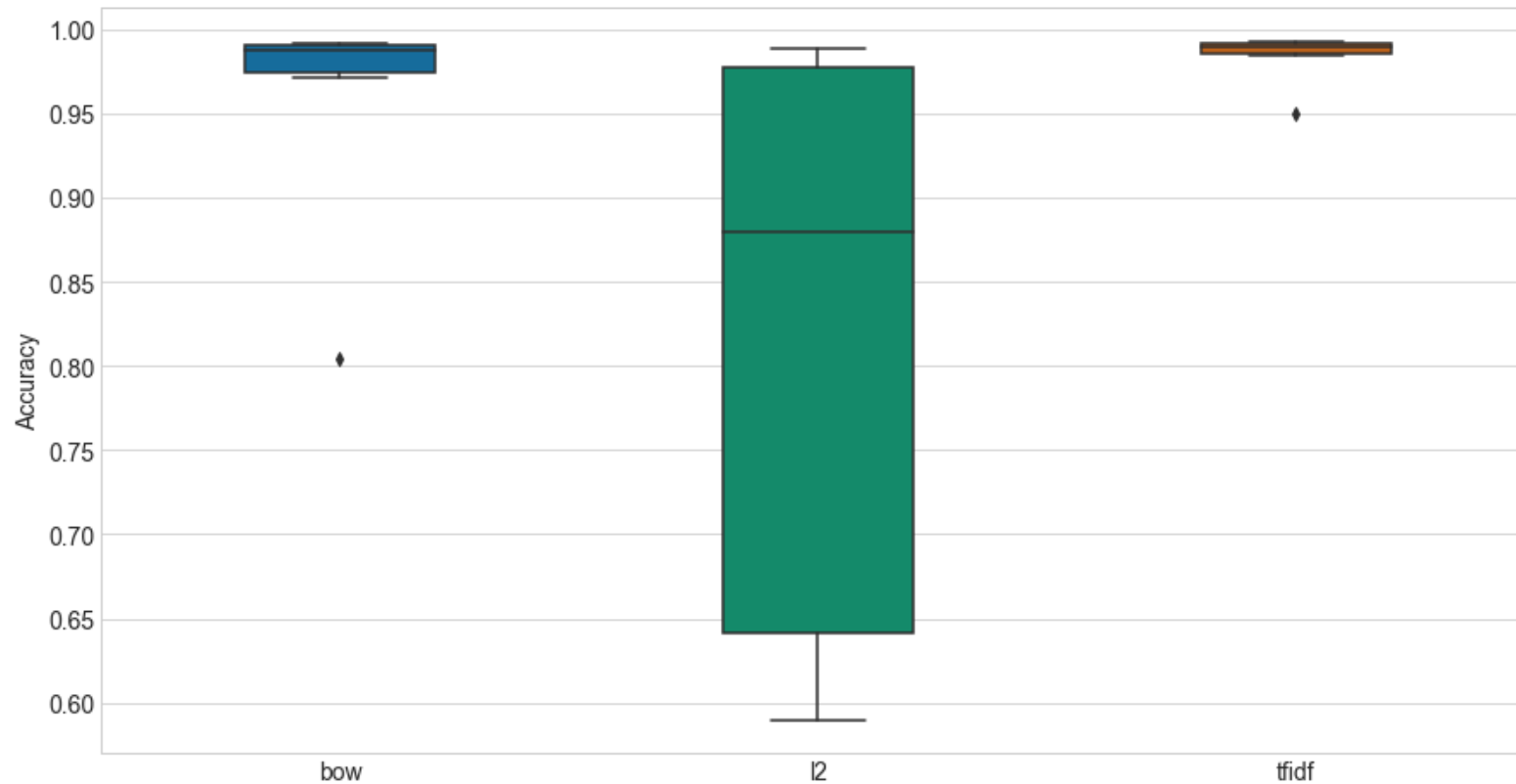
Explanation of Data



After removing Mixed Occupancy values, which are rare occurrences, the Use Type field looks good for binary classification.

Baseline metric = 52.81% accuracy

Feature Engineering

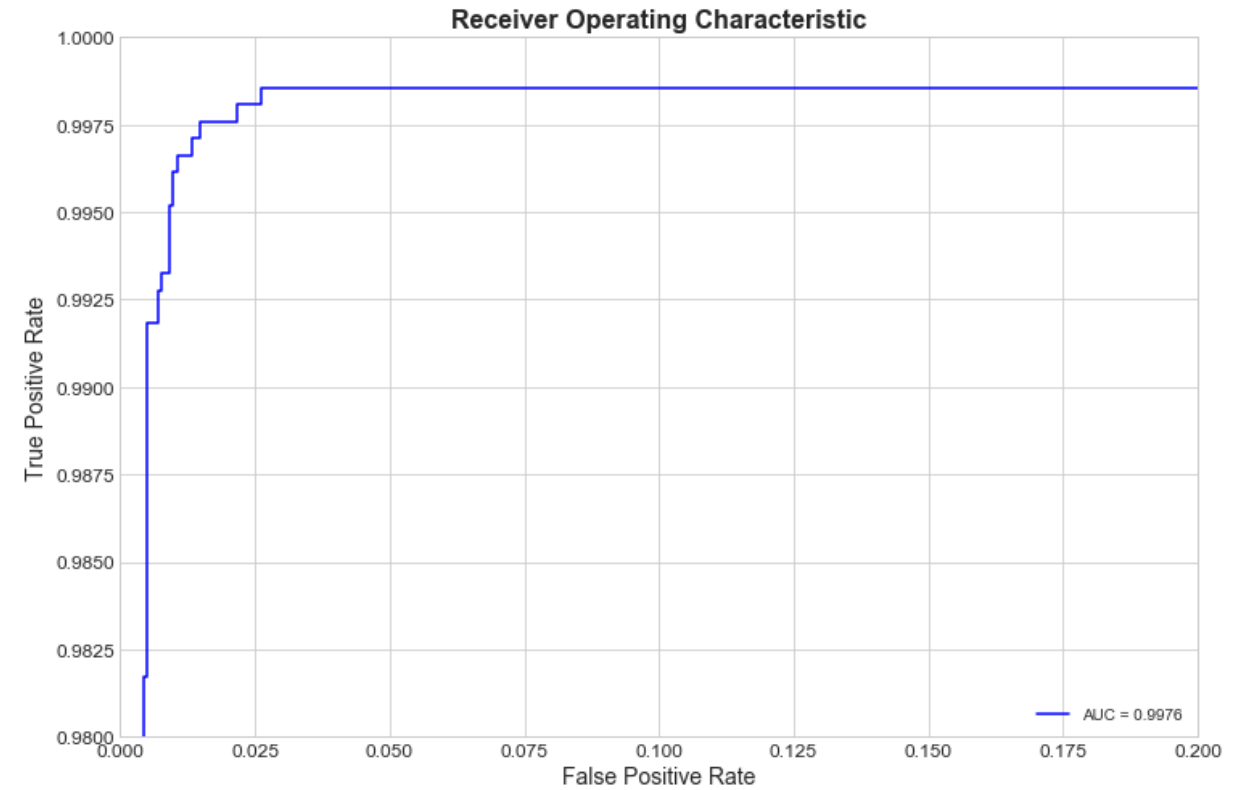
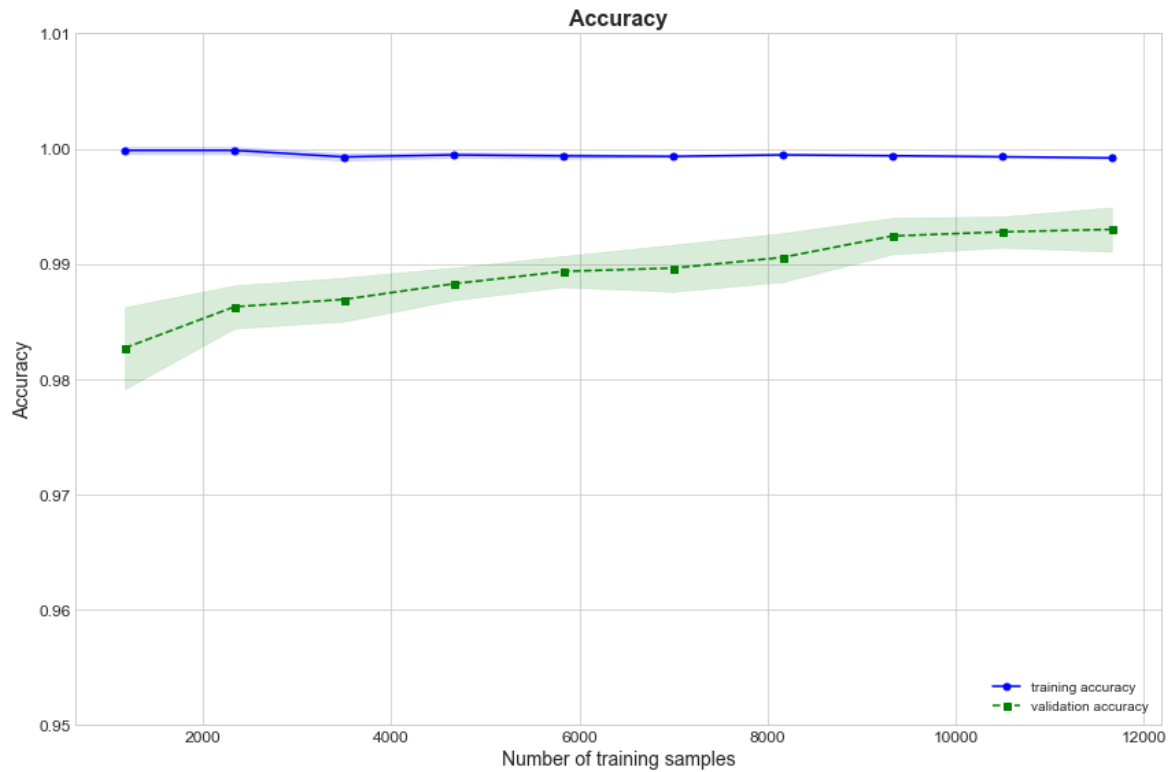


The Description Field was converted to Bag of Words, l2 normalized and tf-idf representations and compared using logistic regression.

The results indicated the tf-idf representation as the top performer.

Model Analysis

■ Logistic Regression



Model Analysis

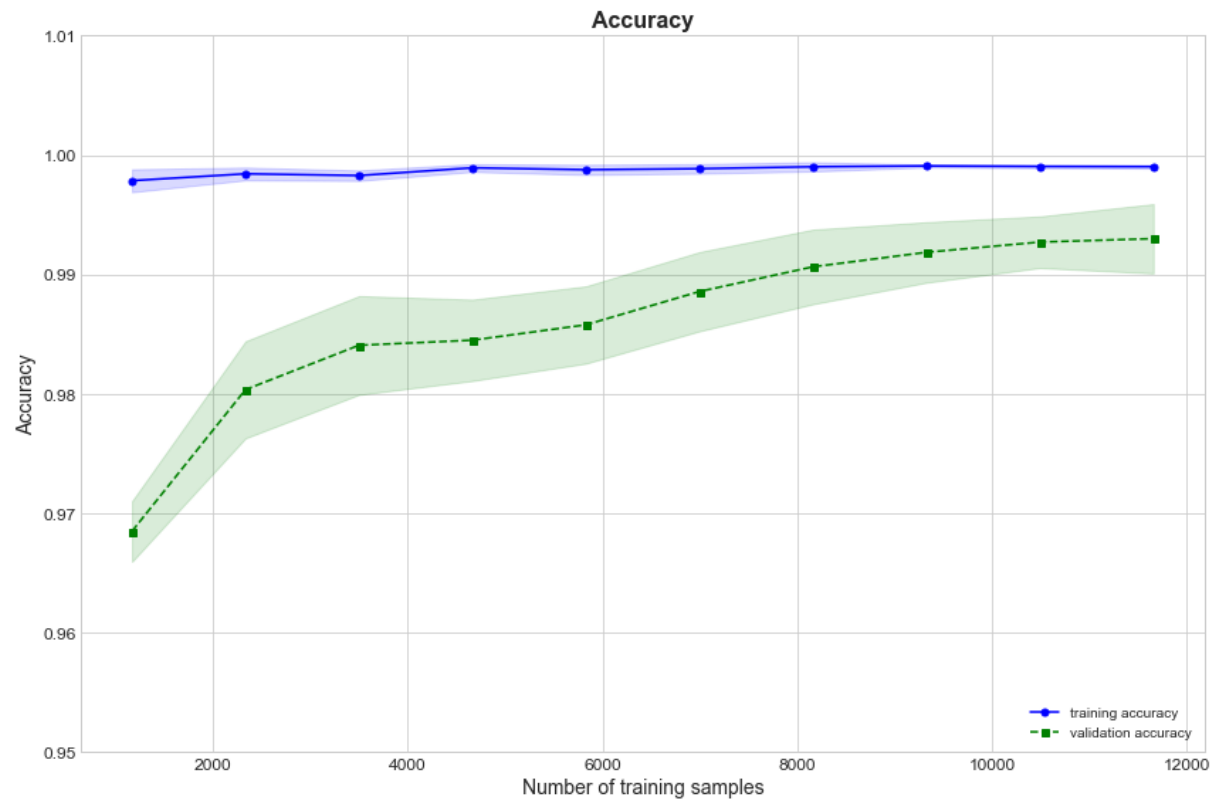
- Logistic Regression

		Commercial	Residential
True label	Commercial	1409	14
	Residential	6	2070
		Predicted label	



Model Analysis

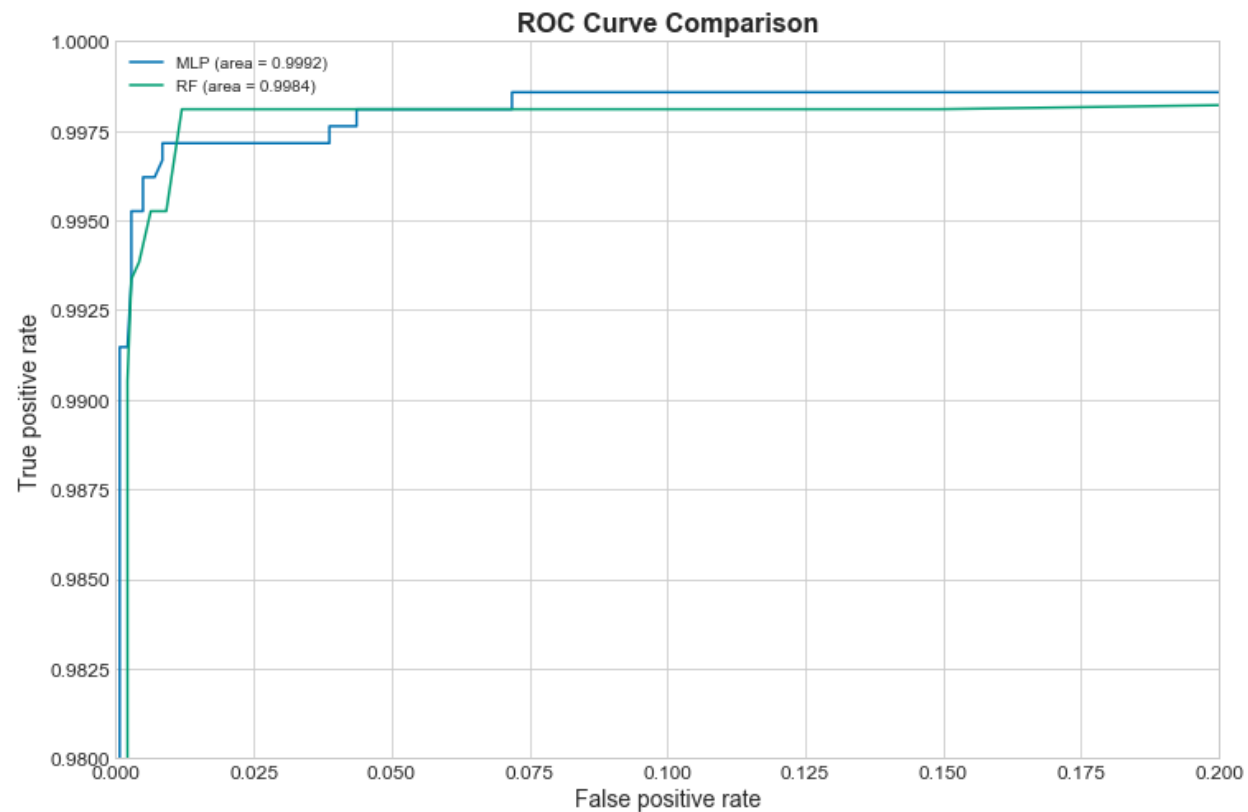
■ Random Forest



		Commercial	Residential
True label	Commercial	1483	12
	Residential	10	1994
		Predicted label	

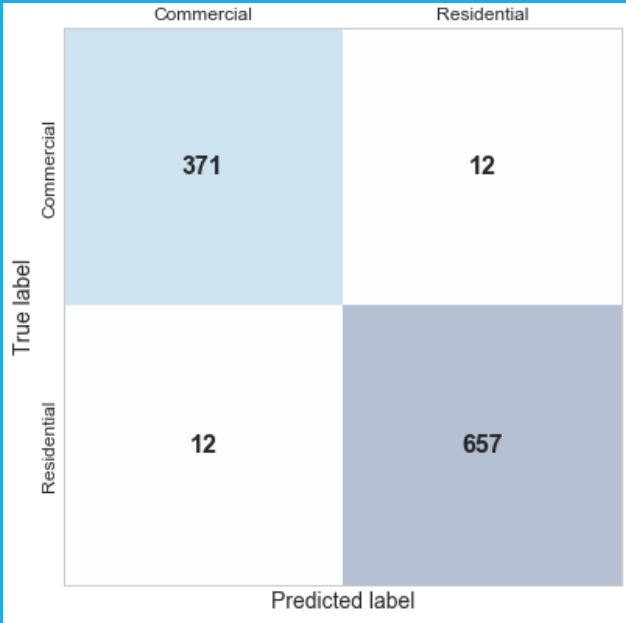
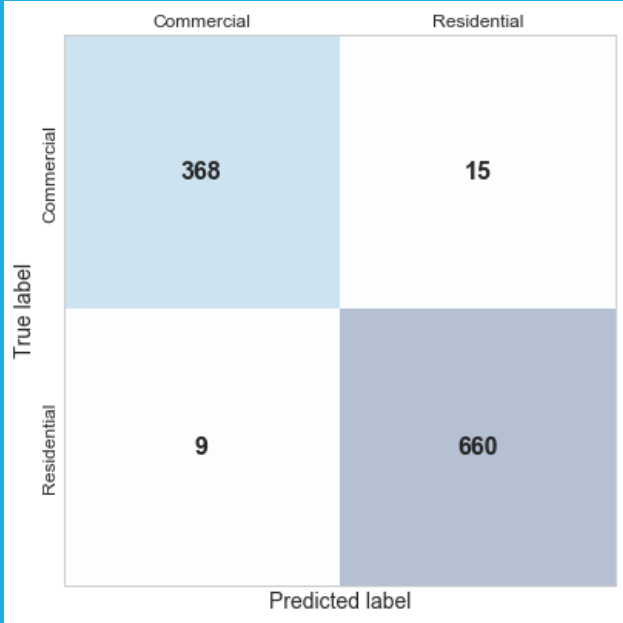
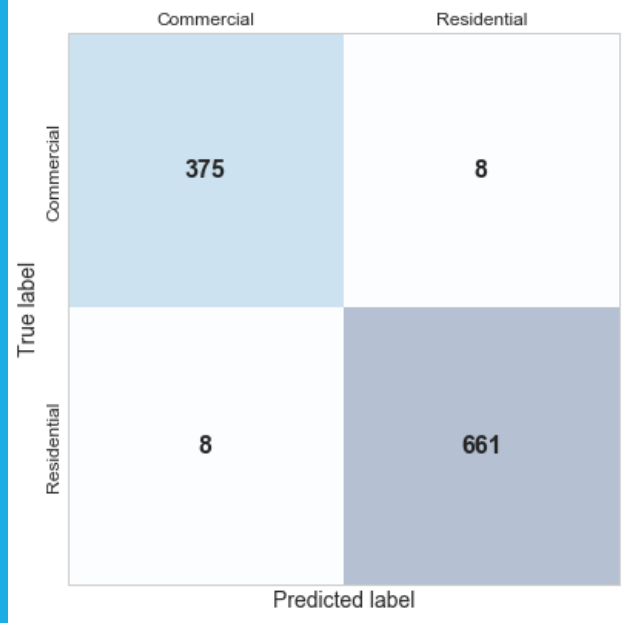
Model Analysis

- Multilayer Perceptron outperforms Random Forest



		Commercial	Residential
True label	Commercial	1483	12
	Residential	11	1993
		Predicted label	

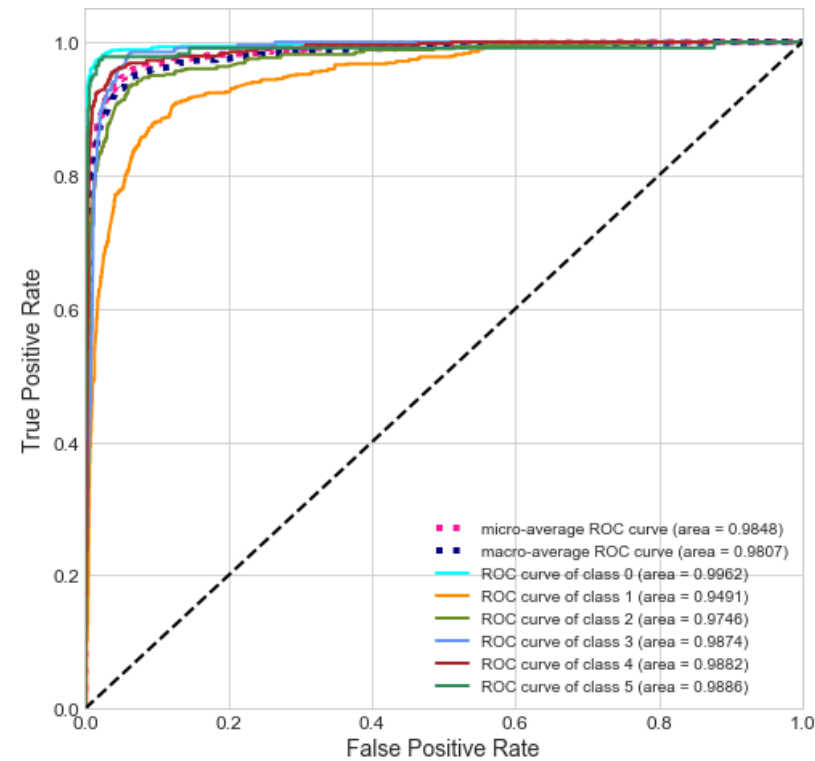
Model Summary

	Logistic Regression	Random Forest	Multilayer Perceptron																											
Test Accuracy	99.34%	97.97%	99.34%																											
Validation Accuracy	97.71%	97.71%	98.47%																											
Confusion Matrix of model on validation set	 <p>A 2x2 confusion matrix for Logistic Regression. The y-axis is labeled 'True label' with categories 'Commercial' and 'Residential'. The x-axis is labeled 'Predicted label' with categories 'Commercial' and 'Residential'. The matrix shows 371 true positives (Commercial predicted Commercial), 12 false positives (Residential predicted Commercial), 12 false negatives (Commercial predicted Residential), and 657 true negatives (Residential predicted Residential).</p> <table><tr><th></th><th>Commercial</th><th>Residential</th></tr><tr><th>Commercial</th><td>371</td><td>12</td></tr><tr><th>Residential</th><td>12</td><td>657</td></tr></table>		Commercial	Residential	Commercial	371	12	Residential	12	657	 <p>A 2x2 confusion matrix for Random Forest. The y-axis is labeled 'True label' with categories 'Commercial' and 'Residential'. The x-axis is labeled 'Predicted label' with categories 'Commercial' and 'Residential'. The matrix shows 368 true positives (Commercial predicted Commercial), 15 false positives (Residential predicted Commercial), 9 false negatives (Commercial predicted Residential), and 660 true negatives (Residential predicted Residential).</p> <table><tr><th></th><th>Commercial</th><th>Residential</th></tr><tr><th>Commercial</th><td>368</td><td>15</td></tr><tr><th>Residential</th><td>9</td><td>660</td></tr></table>		Commercial	Residential	Commercial	368	15	Residential	9	660	 <p>A 2x2 confusion matrix for Multilayer Perceptron. The y-axis is labeled 'True label' with categories 'Commercial' and 'Residential'. The x-axis is labeled 'Predicted label' with categories 'Commercial' and 'Residential'. The matrix shows 375 true positives (Commercial predicted Commercial), 8 false positives (Residential predicted Commercial), 8 false negatives (Commercial predicted Residential), and 661 true negatives (Residential predicted Residential).</p> <table><tr><th></th><th>Commercial</th><th>Residential</th></tr><tr><th>Commercial</th><td>375</td><td>8</td></tr><tr><th>Residential</th><td>8</td><td>661</td></tr></table>		Commercial	Residential	Commercial	375	8	Residential	8	661
		Commercial	Residential																											
Commercial	371	12																												
Residential	12	657																												
	Commercial	Residential																												
Commercial	368	15																												
Residential	9	660																												
	Commercial	Residential																												
Commercial	375	8																												
Residential	8	661																												

Next Steps

Logistic Regression Multiclass Classification | Initial Test Accuracy: 86.04%

True label	Predicted label					
	New	Existing	Alteration	Repair	Other	Addition
New	955	10	2	2	3	1
Existing	76	589	35	38	21	3
Alteration	38	54	488	14	2	2
Repair	20	29	7	408	2	0
Other	20	28	0	1	424	2
Addition	18	2	5	0	0	200



Next Steps

- Continue to fine tune and evaluate models and model performance
- Simplify code and isolate each model into its own notebook
- Save models for quicker analysis
- Further develop models for Job Type multiclass classification
- Create a project proposal and deck for presentation to city stakeholders
- Setup and meet with city stakeholders



Appendix

Data Source:

<https://data.montgomeryal.gov/Permits/Building-Permit-2014-Present-Download-/qvzc-ejq2>

Project Repository:

<https://git.generalassemb.ly/bdub595217/project-final>

Other Sources:

<https://scikit-learn.org>

<https://machinelearningmastery.com>

Zheng, Alice, and Amanda Casari. “Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists.” O'Reilly, 2018.

Géron, Aurélien. “Hands-On Machine Learning with Scikit-Learn and TensorFlow.” O'Reilly, 2017.