Abstract

The construction industry in Ireland, the UK and worldwide is in crisis. Post Covid and with a war raging in Ukraine, there are manpower shortages, price increases and a lack of supply of housing, modern office blocks and infrastructure. The author of this research focused mainly on commencement notices as indicators of construction activity from 2014-2020. Various machine learning models were used for time series forecasting, sentiment analysis and linear regression amongst others. Various statistical tests were employed to test a number of hypotheses. An interactive dashboard was developed to display information in a convenient format. The author concludes by stating that linear regression and logistic regression models worked best. The results of the chi square and Anova statistical tests were interesting also. More research needs to be carried out to address deficits in housing supply in Dublin.

Table of Contents

Introduction – Background – The crisis in construction.

Exploratory Data Analysis

* + - 1. Data collection
      2. Data storage
      3. Filtering the data
      4. Presenting the data

Which ML models do we use and why?

* + - 1. Time series forecasting
      2. Multiple regression
      3. Logistic regression
      4. Sentiment Analysis
      5. Linear discriminant analysis
      6. Principal discriminant analysis
      7. KNN, K-Means and DBSCAN

Introduction

There are many different facets to consider when you look at the construction industry. Building costs, availability of manpower, the price of materials, the location and nature of the buildings, roads, airports that are to be created and, of course, regulation. The author was formerly a public sector official who worked in planning development and control. The main thrust of this assessment will be an investigation of planning commencement notices which are statutory documents that indicate the construction (of a house or office) is about to begin.

In the Civil Service, for example, decisions are made on whether or not to construct a new road or school. In Local Government there is a need for housing/planning authorities to decide upon planning permission(s) for homes, offices and large-scale developments (Google HQ, for instance). Then we rely on builders to construct them and for individuals to reside or work in them. We use data analysis to gather the information for the demand for new constructions. We can then forecast trends for the provision of all manner of buildings, roads and so forth in the coming years.

Planning is an important part of Local Government. Here in Ireland, it is covered by the 1963 Planning, Housing and Development Act (amended in 2000). Basically, if a company like “Google” or an individual submits a planning application for let us say 2 office blocks, once the application has been approved, the applicant submits a commencement notice and within a short time, site building should commence. Usually, delays may be attributed to funding issues, weather, government policy or manpower issues.

Consider this example. Census data has shown a year-on-year growth of the Irish population (<https://www.irishtimes.com/ireland/housing-planning/2022/06/23/>, accessed on 10th April 2023). From this we can safely predict a need for extra homes (ibid.). This is a good example of why we need accurate data in the construction industry.

The primary dataset(s) which the author used contained information on the number of commencement notices submitted yearly throughout Ireland They were selected from [www.data.gov.ie](http://www.data.gov.ie) (accessed on 10th April 2023). The author also availed of datasets from the UK for comparison (ons.gov.uk).

The datasets contained information relating to various matters, such as.

The number of commencement notices submitted yearly per county/ administrative area across Eire, the unit cost of construction (per square metre) in Eire and information upon the types of planning permission granted in Dublin, Ireland and Leicester, England.

## The construction industry in Ireland and abroad. According to an article in the Irish Times the construction industry is experiencing a global labour shortage. Demand for new building projects is high, as markets get back up to speed after a tumultuous year. However, the availability of labour is struggling to return to pre-COVID-19 levels as many countries find the workers who went home to see out the pandemic are yet to return. Combined with a shortage of building materials, it means the wheels of industry are slowly creaking back into action rather than going full steam ahead. The article also mentions that there are various socio-political causes of labour shortages in different countries.

“British construction firms have been poaching Irish-based construction workers” who lost their jobs during Covid (2020-2021). This was part of an alarming headline from a few years ago when there were fears that our (Irish) construction sector would witness that the “flow of builders across the Irish Sea [would] become a permanent skill drain and [would] damage the industry” long-term (ibid.). When many commercial and industrial construction sites remained closed, a number of UK construction recruitment firms were actively seeking Irish-based workers. This “led to a growing number of builders taking jobs elsewhere”. Sources like Flynn Construction, which focuses on commercial projects such as data centres and healthcare facilities, say they were “directly aware of a growing number of firms who had lost skilled tradespeople that would be difficult to replace”.

Building slowed at its fastest rate since the summer last month as inflation continued to hit the industry. The rise in energy costs was a big factor.

The latest construction survey shows that building slowed for the third month running in December, with rising costs hitting demand (Q4 of 2022).

The Construction Purchasing Managers’ Index, seen as a key indicator of industry trends, slid to 43.2 in December, its sharpest fall since July. To put this in context “any return below the index’s benchmark of 50 means activity shrank, while any reading above that number indicates that it expanded”. The article continued.

“However, we expect to see further strong activity in the logistics sector as occupier demand for warehousing space has led to low vacancy rates and solid rental growth.”

Meanwhile, figures released by the Department of Housing recently show the number of residential units where construction has commenced had fallen 10 per cent on the year to 27,417 in September on a 12-month rolling basis – down from 34,846 in March and driven by a fall-off in apartment starts.

“With demand remaining muted, construction firms scaled back their purchasing activity for the fifth month running, and to a greater extent than in the previous survey period,” BNP Paribas Real Estate said.

## Brexit and the UK construction industry. Another factor that has had an impact on both the Irish and British construction sector has been Brexit.

## There were 5 actions recommended that UK companies of all sizes should take:

## Analyse your existing supply chains to identify key risks as well as potential competitive opportunities. A key risk area would be understanding if subcontractors are Brexit ready.

## Understand your competitor landscape and ways to gain advantage through repositioning your organisation for growth.

## Engage with key customers to understand their specific Brexit related needs and ways in which you can support them, be it through new housing developments for staff moving to Ireland, more office space or general renovations to deal with the increased space demands.

## Consider the need for contingency planning which will minimise any supply chain delays on your business. Ensure all necessary customs registrations and a customs broker is in place. You should ensure that you are satisfied that all goods continue to meet the necessary safety and regulatory frameworks within the EU. If not, then consider looking at alternative supply chain routes from within the EU to source your materials. Furthermore, companies should engage with all Brexit related supports currently available through various Government agencies.

## Undertake financial models (and, by implication, Machine Learning modelling) to understand the impact of Brexit on your business. Consider areas such as potential sterling devaluation, increased costs driven by tariffs and new customs requirements, and the associated cash flow implications.

## Clearly, the effects of Brexit and the war in Ukraine have affected construction costs.

## One other factor has to be addressed. Irish planning laws are a mess. Anyone who makes an observation or raises an objection to a proposed construction can appeal the decision of a local authority and delay it for months or years. In some case the Appeals board overturn decisions and refuse planning permission outright. Unfortunately NIMBY-ism is rife.

How can ML modelling help us? Firstly lets define our terms.

**Artificial Intelligence (AI)**, is “the ability of a digital [computer](https://www.britannica.com/technology/computer) to perform tasks commonly associated with intelligent beings” ([www.Britannica.com](http://www.Britannica.com), accessed on 1st April 2023) . Machine1 learning (ML) “is considered a subset of Artificial Intelligence because it enables the extraction of meaningful patterns from samples, which is a capability of human intelligence” ([www.medium.com](http://www.medium.com), accessed on 10th April 2023).

We can therefore state **Machine learning** is**a subset of “artificial intelligence”, which is defined as the capability of a machine “to imitate intelligent human behaviour”** (Lao, R. 2018; Tan, O. 2017; Wolfiewicz, W. 2022). It has much to offer in this particular area of study, namely, construction issues.

ML is ideal for classification problems, e.g. “Will there be enough houses bult in Ireland before 2030 or not?” and regression problems, for example, “If Meath County Council received 568 commencement notices this year, how many will they receive next year? This represents Supervised Learning.

These are some techniques that are commonly used in Unsupervised Learning such as “nearest-neighbour mapping, and k-means clustering” which may prove beneficial to this assignment as their main goal is to explore the data and find some structure within.

## The author used a data mining “roadmap” ([www.machinelearningmastery.com](http://www.machinelearningmastery.com)) to assist his endeavours, where.

1. Data Collection: “The quantity and quality of our data will directly specify how good” our predictive model can be ibid.). The data mostly was derived from [www.data.gov.ie](http://www.data.gov.ie) and ons.gov.uk.

2. Data Preparation: The data needed to be prepared, normalized, checked for duplicates and errors removed. Visualization of the data was used as a technique to find patterns and outliers to see if the required information had been collected properly.

3. Choose a Model: There are a number of different models for various tasks and goals. The author chose (at least one) classification, clustering, time series and sentiment analysis model.

4. Train the Model: The objective of training is to answer a question or make a prediction correctly as often as possible. This means it uses your training data and incrementally enhance the predictions of the model. Each cycle of updating the weights and biases is considered as one training step.

5. Evaluate the Model**:**Some metric or combination of metrics is required to measure objective performance of model are used in this step. It means you should test the model against previously unseen data to see how it performs.

6. Parameter Tuning: In other words, *you* should set parameters to improve the process. In KNN you might try k=3 or k=2, or the elbow method to find the optimal number of nearest neighbours.

7. Make Predictions: Machine Learning “uses data to answer questions. Thus, Prediction or inference, is the final step to where we get to answer some questions” (ibid.).

## One final point to note is we used Grid Search CV to optimise our models. GridSearchCV is the “process of performing hyperparameter tuning in order to determine the optimal values for a given model” (<https://www.mygreatlearning.com/blog/gridsearchcv/>, accessed on 10th April 2023).

## ML models can help us to predict trends (e.g. how to forecast how many houses are required, how many schools and roads are required), they can classify various issues for us, such as, (will the development constitute a health hazard?, is it fit for purpose?, is it over-budget) and they can look for clusters of hidden information amongst all the data.

## Exploratory Data Analysis:

## The author chose appropriate datasets from the UK and Ireland.

## Using python, these .csv and .json files were wrangled into an acceptable format. Redundant columns were deleted, and important features were identified.

## Hypotheses were created and investigations were carried out with ML tools.

## Both static and interactive dashboards were used to present the results visually,

Conclusion:

The entire point of this exercise was to analyse a dataset of reports that relate to construction in Ireland and abroad to look for insights into the housing crisis.

We made a number of hypotheses, investigated various correlations and differences in the data, applied some ML models and formulated our results.

### We found that Chi square was “W”. Hypothesis: H0 states the presence (or absence) of Activity on Site is **independent** of whichever local authority is in charge. H0 was rejected.

Student t-test. H0 states that mean values of the columns (Units Completed to Date) is no different under the criteria "Activity on Site" or no "Activity on Site". We found insufficient evidence to reject H0.

Logistic regression - classification (where the predicted variable was the Local Authority, accuracy was 0.694).

Logistic regression - classification (where the predicted variable was the presence of Activity on Site, accuracy was 0.9915).

Multiple regression Our model predicted 105 units (permitted but not commenced) based on our predictor variables.

Logistic regression (for regression), accuracy was 0.985.

Random Forest – classification Our model had an accuracy of 1.0. This can be attributed to over fitting.

Random Forest - regression (Our RMSE was 46.482). This is very high. It suggests over reliance on a particular variable.

KNN - classification (k=3 was optimal, accuracy was 0.93877).

#### GridSearchCV was used to optimize our classifier from accuracy of 0.706 to 0.744 (l1 regularization, lasso regression).

#### GridSearchCV was used also to optimize our regression model (k=5).

Simple linear regression (r squared was 0.738) between predictor and target variables.

Naïve Bayes Accuracy score was 0.69421.

In conclusion, data science has much to offer in the understanding and interpretation of the housing reports gathered by the task force. Worryingly, the data shows that many of the projects granted permission in the first quarter of 2018 had either not commenced construction or where there had been construction, no activity was taking place on site. The consistent supply of houses/apartments for the people of Dublin and elsewhere is self-evidently important and deficits in the construction of homes (identified by data analysis) must be addressed.

References

Ashok, P. (2022). What is regression? Definition of regression. Available at: <https://www.mygreatlearning.com/blog/what-is-regression/> (accessed on 10th March 2023).

Carswell, S. (2023). Population increase continues to outstrip growth in number of new homes. Available at: <https://www.irishtimes.com/ireland/housing-planning/2022/06/23/census-2022-population-increase-continues-to-outstrip-growth-in-number-of-new-homes/> (accessed on 10th April 2023).

Copeland, B.J. (2023). Artificial Intelligence. Available at: <https://www.britannica.com/technology/artificial-intelligence> (accessed on 10th March 2023).

Copeland. E et alia. (2018). Setting up an office of data analytics. Available at: [www.nesta.org/](http://www.nesta.org/) (accessed on 10th March 2023).

## DHTF Data Q1 2018 - ArcGIS. (2023). Available at: https://[www.data.gov.ie](http://www.data.gov.ie) (accessed on 10th March 2023).

Dublin City Council. (2023). Available at: <https://www.dublincity.ie/residential> (accessed on 23rd March 2023).

# Finegan, N. 2023. New online e-Planning system to modernise application and submission process. Available at: <https://www.meathchronicle.ie/2023/02/14/new-online-e-planning-system-to-modernise-application-and-submission-process/> (accessed on 10th April 2023).

# Geeks for Geeks. (2023). DBSCAN Clustering in ML. Available at: <https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/> (accessed on 10th April 2023).

Geeks for Geeks. (2023). Logistic regression in Machine Learning. Available at: <https://www.geeksforgeeks.org/understanding-logistic-regression/> (accessed on 10th April 2023).

# Geeks for Geeks. (2023) K means Clustering. Available at: <https://www.geeksforgeeks.org/k-means-clustering-introduction/> (accessed on 10th April 2023).

Gandhi, R. (2021). Naïve Bayers classifier. Available at: <https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c>/ (accessed on 10th April 2023).

# Geeks for Geeks. (2023). DBSCAN Clustering in ML. Available at: <https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/> (accessed on 10th April 2023).

Geeks for Geeks. (2023). Logistic regression in Machine Learning. Available: <https://www.geeksforgeeks.org/understanding-logistic-regression/> (accessed on 10th April 2023).

Ghia, A. et alia. 2023. Accelerating data and analytics transformations in the public sector. Available at: <https://www.mckinsey.com/industries/public-and-social-sector/our-insights/accelerating-data-and-analytics-transformations-in-the-public-sector> (accessed on 10th March 2023).

Government Publications. Planning and Development Act 2000. Available at: <https://www.irishstatutebook.ie/eli/2000/act/30/enacted/en/html/> (accessed on 20th March 2023).

Government Publications (2014). Available at: <https://merrionstreet.ie/en/wp-content/uploads/2014/05/construction-strategy-14-may-20141.pdf> (accessed on 17th March 2023).

Government Publications. (2023). Available at: [www.opendata.housing.gov.ie](http://www.opendata.housing.gov.ie) (accessed on 10th March 2023).

Great Learning Team. (2022). What is GridSearchCV. Available at: <https://www.mygreatlearning.com/blog/gridsearchcv/> (accessed on 10th April 2023).

IBM. (2023). K Nearest Neighbour. Available at: <https://www.ibm.com/docs/en/ias?topic=procedures-k-nearest-neighbors-knn/> (accessed on 10th April 2023).

IBM. (2023) What is Supervised learning? Available at: <https://www.ibm.com/topics/supervised-learning> (accessed on 10th March 2023).

IBM. (2023) What is Unsupervised learning? Available at: <https://www.ibm.com/topics/supervised-learning> (accessed on 10th March 2023).

# Javatpoint Team. (2023). K-Nearest Neighbour (KNN). Available at: Algorithm for Machine Learning. <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning/> (accessed on 10th April 2023).

Kaseb, Kaynan (2020) Understand Machine Learning. Available at <https://medium.com/kayvan-kaseb/understand-machine-learning-e140fc93bdf1> (accessed on 19th March 2023).

Lao, R. (2018). [A Beginner’s Guide to Machine Learning](https://medium.com/@randylaosat/a-beginners-guide-to-machine-learning-dfadc19f6caf). Available at: <https://medium.com/@randylaosat/a-beginners-guide-to-machine-learning-dfadc19f6caf> (accessed on 10th March 2023).

Leinhardt, G., Leinhardt, S. (1980). [Exploratory Data Analysis: New Tools for the Analysis of Empirical Data](https://journals.sagepub.com/doi/pdf/10.3102/0091732X008001085), Review of Research in E0ducation, Vol. 8, (1980), pp. 85–157.

Mitchell, Tom M. (1997) Machine Learning. USA. McGraw Hill.

Mitchell, Tom M. (2006). The discipline of machine learning. USA. School of Computer Science, Carnegie Mellon University.

# Patwardhan, S. (2021). Simple understanding and implementation of KNN algorithm. Available at: <https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-of-knn-algorithm/> (accessed on 1st April 2023).

Statology. (2023). Introduction to logistic regression. Available at: <https://www.statology.org/logistic-regression/> (accessed on 10th April 2023).

Starmer, Josh (2018) Logistic Regression. Available at: <https://statquest.org/statquest-logistic-regression/> (accessed on 19th March 2023).

Tan, O. (2017) How does a Machine Learn? Available at: <https://www.forbes.com/sites/forbestechcouncil/2017/05/02/how-does-a-machine-learn/?sh=208637ac744/> (accessed on 19th March 2023).

[The Complete Guide to Machine Learning Steps](https://d.docs.live.net/f5281318a397632c/Documents/The%20Complete%20Guide%20to%20Machine%20Learning%20Steps) (2020). Available at: <https://www.simplilearn.com/tutorials/machine-learning-tutorial/machine-learning-steps> (accessed on 19th March 2023).

Tukey, John W. (1977). [*Exploratory Data Analysis*](https://en.wikipedia.org/wiki/Exploratory_Data_Analysis). New York. Pearson.

# Van Neestra, Anne Fleur et alia (2020). The use of public sector data analytics in the Netherlands. Available at: <https://www.emerald.com/insight/content/doi/10.1108/TG-09-2019-0095/full/html> (accessed on 10th March 2023).

Walker, S. and Duncan, D. (1967). "Estimation of the probability of an event as a function of several independent variables". Biometrika. 54 (1/2): 167–178).

What is Supervised Learning? (2020). Available at: <https://www.ibm.com/cloud/learn/supervised-learning> (accessed on 21st March 2023).

What is Unsupervised Learning? (2020). Available at: <https://www.ibm.com/cloud/learn/supervised-learning> (accessed on 21/03/2023).

What is AI? Learn about Artificial Intelligence. (2022) Available at: <https://www.oracle.com/ie/artificial-intelligence/what-is-ai/> (accessed on 21st March 2023).

[Wolfewicz, Arne (2022). How do machines learn? A beginner’s guide. Available at: https://levity.ai/blog/how-do-machines-learn/ (accessed on 21](https://d.docs.live.net/f5281318a397632c/Documents/Wolfewicz, Arne (2022). How do machines learn?%20A%20beginner’s%20guide.%20%20Available%20at:%20https://levity.ai/blog/how-do-machines-learn/%20%20(accessed%20on%2021st%20March%20/2023).)[st](https://d.docs.live.net/f5281318a397632c/Documents/Wolfewicz, Arne (2022). How do machines learn?%20A%20beginner’s%20guide.%20%20Available%20at:%20https://levity.ai/blog/how-do-machines-learn/%20%20(accessed%20on%2021st%20March%20/2023).) [March /2023).](https://d.docs.live.net/f5281318a397632c/Documents/Wolfewicz, Arne (2022). How do machines learn?%20A%20beginner’s%20guide.%20%20Available%20at:%20https://levity.ai/blog/how-do-machines-learn/%20%20(accessed%20on%2021st%20March%20/2023).)

Yufeng, G. (2017). The 7 steps of Machine Learning. Available at: <https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e> (accessed on 10th March 2023).

Appendix

\*\*Data Dictionary\*\*

Object ID - Identifier of file.

Planning Authority - Identifies Dublin City Council, Fingal, SDCC or Dun Laoighaire.

Tier 1 - Denotes importance of planning application (e.g., 1 is housing/apartments, 2 is private dwellings, 3 is extension/conversion et cetera)

Planning Reference - The last 2 digits show the year that an application was lodged, first few digits denote the order of application (p.p.25/14 means that the application was the 25th application of 2014, for example).

ITM\_X, ITM\_Y Coordinates (Irish Transverse Mercator (ITM) is the geographical coordinate system for Ireland).

Planning\_Permission\_\_Units\_Perm - Total no. of planning permssions granted.

Units\_Completed\_to\_Date - Total no. of housing units finished to date.

Units-under-Construction - Total no. of housing units where work has started but not completed.

Activity\_On\_Site - A boolean style value. Y shows there is work being done on site, N shows no work being done.

Units\_Permitted\_But\_Not\_Commenced - Planning permission has been given but no work has begun.

Planning\_Search\_URL - The website address where one can view the application.