Coursework

7CS030/UM1: Concept & Technologies of Artificial Intelligence

Faraz Yusuf Khan, 2307375

SCHOOL OF MATHEMATICS AND COMPUTER SCIENCE, FACULTY OF SCIENCE AND ENGINEERING, THE UNIVERSITY OF WOLVERHAMPTON

Contents

Task 1: Regression

1.0 Introduction

This section is dedicated to real estate analysis using Linear Regression techniques to forecast house prices in King County, USA. The provided dataset consists of 18 features with features such as floor levels, living area in square feet and number of bathrooms.

1.1 Model Description and Algorithm

Starting with a simple model, we deploy Simple Linear Regression where we predict the house process on the basis of the square footage of living space (sqft_living). Linear Regression is utilized for modeling the relationship between the predictor variable(s) and a target variable, it uses a linear equation on the observed data (Maulud and Abdulazeez, 2020). It should be noted that Linear Regression assumes a linear association between target and predictor models.

Predictor features are selected on the basis of their correlation coefficient with House Price and thus, *sqft_living* is selected. Initial analysis of the scatterplot evidences a positive linear relationship between our variables. The data is split into 33% for the testing phase and the rest for analysis.

With Multiple Linear Regression, initially preprocessing of the input features is performed with outliers being visually identified and removed. Additionally, features are tested for skewness and remedied accordingly.

1.2 Results

Model Deployed	Regression Coefficient	Coefficient of determination
Linear Regression	273.9784251	0.5
Multiple Linear Regression	-7012.13, 294.48, -21,0.48	0.48

Table 1.1: Model Metrics

Table 1.1 represents the model metrics for Simple Linear Regression and Multiple Linear Regression respectively. The regression coefficient for Linear Regression implies that for every unit of change in *sqft_living* the House Price will rise by almost 274 units. Our Simple Linear Regression tells us that 50% of the variation in House Price will be due to square feet of living area. Figure 1.1 illustrates the testing and training model performance for our Simple Linear Regression model. It can be noticed that the data points are located far away from the generated line of best fit, further justifying the coefficient of determination at 0.50.

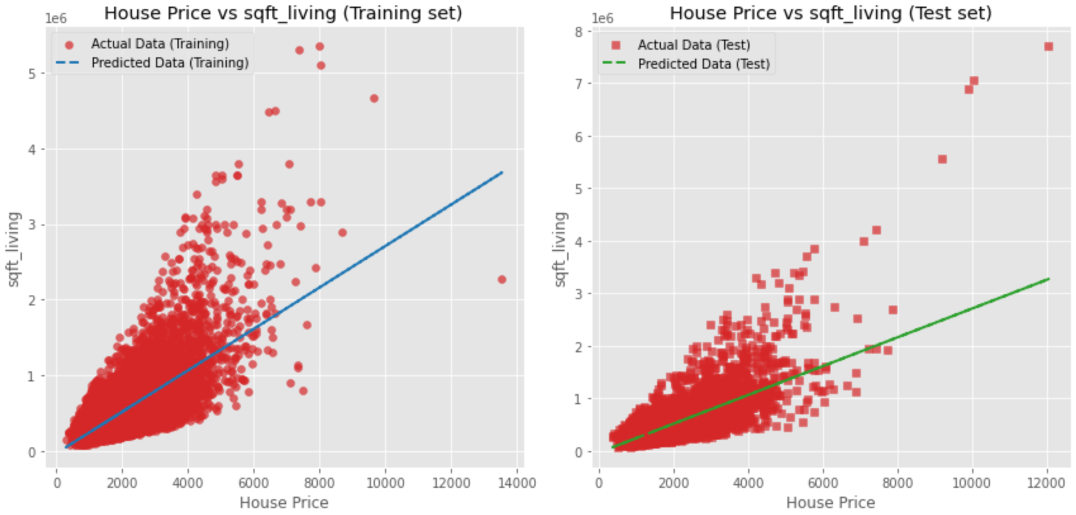

Figure 1.1: Training and Testing Model Performance for Simple Linear Regression

For, Multiple Linear Regression the House Price is lowered by 7012 units and 21 units for the number of bathrooms and living areas above. Furthermore, all three input features explain 48% of the variation in House Price. A visual representation of actual data and predicted data for our Multiple Linear Regression model is illustrated in Figure 1.2, additionally a color-blind friendly visualization depicting the same has also been generated. It can be noticed that the predicted data does not fully match the actual data. Thereby, confirming the coefficient of determination at 0.48.

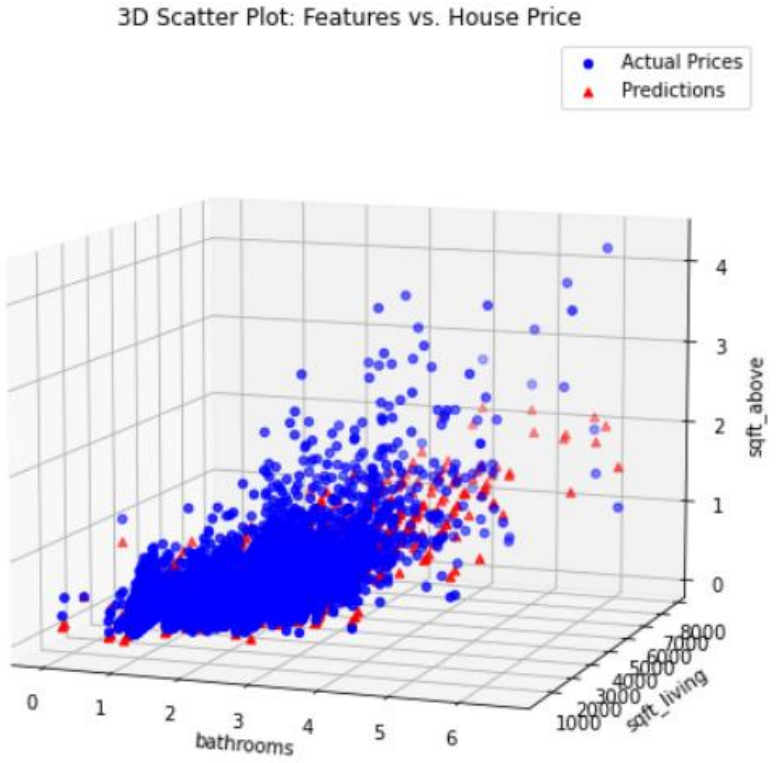

Figure 1.2: Actual Data and Predicted Data for Multiple Linear Regression

1.3 Further Improvements

Both the models give a similar performance as evidenced by the coefficient of determination which only explains a maximum of 50% of variation in house price due to selected predictor variables. Hyperparameter Tuning is recommended as manually trying for different combinations of parameter values will be time consuming. It is suggested that a test for multicollinearity is performed on the variables. Additionally, residual analysis should be performed so that it can be understood where the model is struggling. Furthermore, the presence of overfitting in the training model should be investigated as it may be a case where noise has been included in place of underlying patterns.

1.4 Recommendations

It can be concluded that both models explain a maximum of 50% variability in House Prices which can be considered moderate. A square foot of living area does increase the house price however number of bathrooms and living areas above have been found to lower the house price. Therefore, it is recommended that further analysis consisting of features such as location, neighborhood, age of property, and amenities be performed. Furthermore, consulting with local real estate experts who can provide nuanced analysis on what factors affect the House Price for a certain area.

Additionally, data has evidenced that bathrooms drive down the price. Thus, it is recommended that their conditions be checked and scrutinized and if found sub-par they be put under renovation. Lastly, market trends need to be analyzed as more square feet above did not translate into an increase in House Price. Considering the aforementioned statement, it is imperative that the current economic and market situation be considered.

Task 2: Clustering

2.1 Introduction

This section is dedicated to applying the K-Means Clustering Algorithm to a country dataset consisting of 10 features such as life expectancy, child mortality, and income magnitude. Initially, a simple model consisting of two features will be made followed by a complex model consisting of data preprocessing procedures and all relevant features.

2.2 Model Description and Algorithm

K-Means Clustering will be utilized for making models, it initially guesses the positions of centroids for each cluster and assigns each data point to the nearest centroid. Furthermore, the centroids are recalculated as the mean of all data points within their respective clusters. This step repositions the centroids to the centers of the clusters. The assignment of centroids and their consequent updating continue until a criterion such as the maximum number of iterations is met (de Souza and Costa, 2022).

For our Simple K-Means Clustering Model, the features of income and GDP have been selected. This has been on the basis of a correlation analysis between all variables. These data points will be used to construct two clusters. Our simple model will run a maximum of 50 times to ensure that an optimum location of centroids is found.

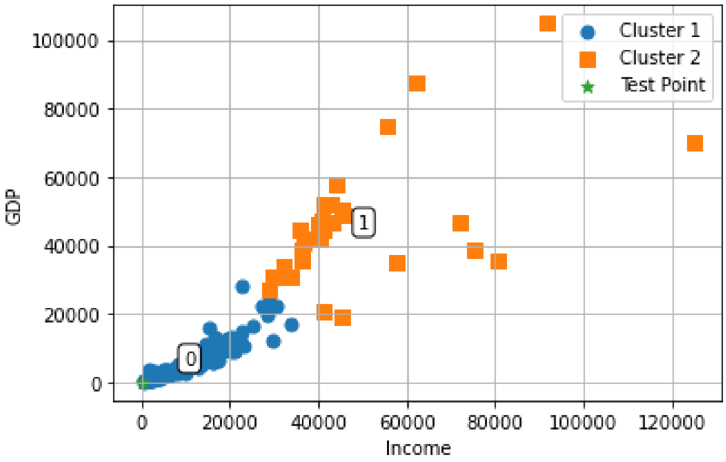

Figure 2.1: Clustering for Simple K-Means Model

Figure 2.1 is dedicated to displaying the modelled clusters based on Income and GDP, we can discern that there are two sets of countries based on our model namely, Developed Countries and Non-Developed Countries. Cluster 1, belongs to Non-Developed Countries and Cluster 2 belongs to Developed Countries based on GDP and Income, respectively.

For, the complex K-Means Clustering Model, the dataset has been modified as it was observed that exports, health, and imports were given as a percentage of GDP, they have been subsequently converted, and correlation analysis has been performed. Outliers have been identified with the help of boxplots and removed with flooring and ceiling operations since there was a presence of numerous outliers and removing them altogether would lead to insufficient data for analysis. Hopkins Statistic is found to lie between 0.84 and 0.91. Therefore, based on the Hopkins statistic magnitude the given dataset has a high tendency of clustering.

The dataset is then scaled to prevent the dominance of certain features over others. Furthermore, a number of three clusters is selected on the basis of an elbow graph representing Inertia (Within Cluster Sum of Squares) and a number of clusters. The number of iterations has been set to 500 with a fixed random state to ensure the reproducibility of results. The generated clusters from our complex K-Means model have been illustrated in Figure 2.2.

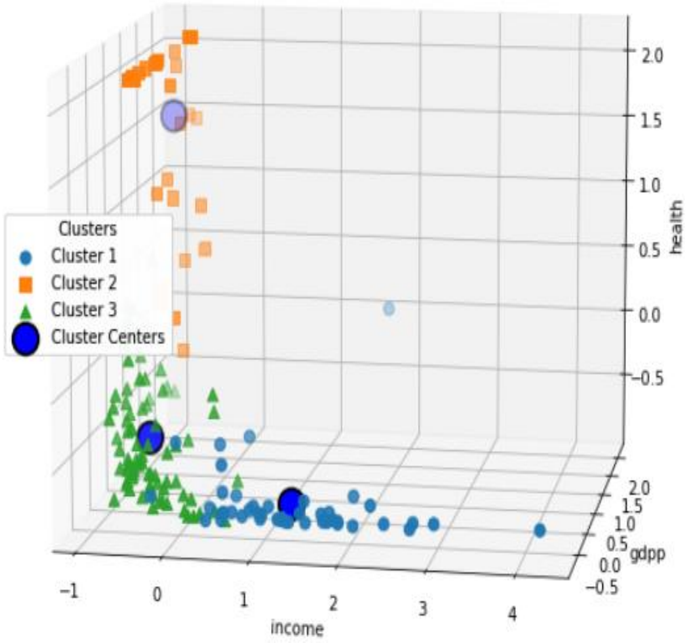

Figure 2.2: Clustering for complex K-Means Model

Figure 2.2 reveals that the clusters can be identified based on Underdeveloped, Developing, and Developed Countries, these classifications are made based on their income and GDP levels. Developed countries have the highest income and GDP magnitude followed by Developing Countries and lastly positioned are underdeveloped countries. The countries were identified and mentioned separately.

2.3 Further Improvements

The dataset can be appended to include more countries for a comprehensive analysis. Additionally, Silhouette scores should be generated so that the best clustering solution can be implemented. Furthermore, it is suggested that features such as literacy rate, educational expenditure, poverty rate, and Internet Penetration Rate be included.

2.4 Recommendation

The Complex K-Means Clustering model is deemed more suitable for selection as it divides the dataset in a more detailed manner. We would recommend aid allocation on the basis of countries identified, with Underdeveloped countries deserving the most magnitude of aid. Health and education should be tailored to the needs of underdeveloped countries focusing on basic education and infrastructure facilities.

For investors, developing countries can be termed as a moderate and safe haven for investment while underdeveloped countries offer a potentially better rate of investment, albeit with a greater risk.

Countries identified as Developed have the potential for cooperation, establishment of trade alliances, and sharing of knowledge with other advanced countries. It is recommended that programs that promote collaboration, establish trade associations, and information exchange between developed nations be encouraged. This could entail helping organizations or businesses in these countries form alliances, joint ventures, or technology exchange initiatives.

Task 3: Classification and Neural Networks

3.1 Model Description and Algorithms Used

Simple models utilize points per game and field goal attempts as predictors, while more complex models use the entire dataset. Logistic Regression (LR) is chosen due to its binary nature. LR assigns probabilities to whether a player lasts 5 years in the NBA. It uses a linear combination of features with a logistic function and a threshold for classification (Huang, 2019). The Artificial Neural Network, a Multi-Layer Perceptron (MLP) classifier, is deployed with two hidden layers of 10 nodes, using logistic activation and 2000 iterations (Cinar, 2020). Gaussian Naïve Bayes assumes feature independence, relying on a normal distribution. Probabilities are calculated based on feature distribution properties, combined with class probabilities (Ismail and Reza, 2022).

3.2 Simple Model Performance

Model deployed	Accuracy (%)	Mislabelled Points	Total Points
Artificial Neural Network	71.05	77	266
Gaussian Naïve Bayes	59.02	531	1329
Logistic Regression	68	503	1329

Table 3.1: Model Performance

Table 3.1 displays the model performance of the Logistic Regression, Gaussian Naïve Bayes, and Artificial Neural Network (ANN) respectively. An accuracy of 71% was calculated for ANN with 77 mislabeled points out of 266 total points represented in Figure 3. On the other hand, Gaussian Naïve Bayes (GNB) had an accuracy of approximately 60% with 531 mislabeled data points out of 1329 total data points. Lastly, Logistic Regression had an accurate prediction rate of 68% with 503 mislabeled data points out of 1329 points.

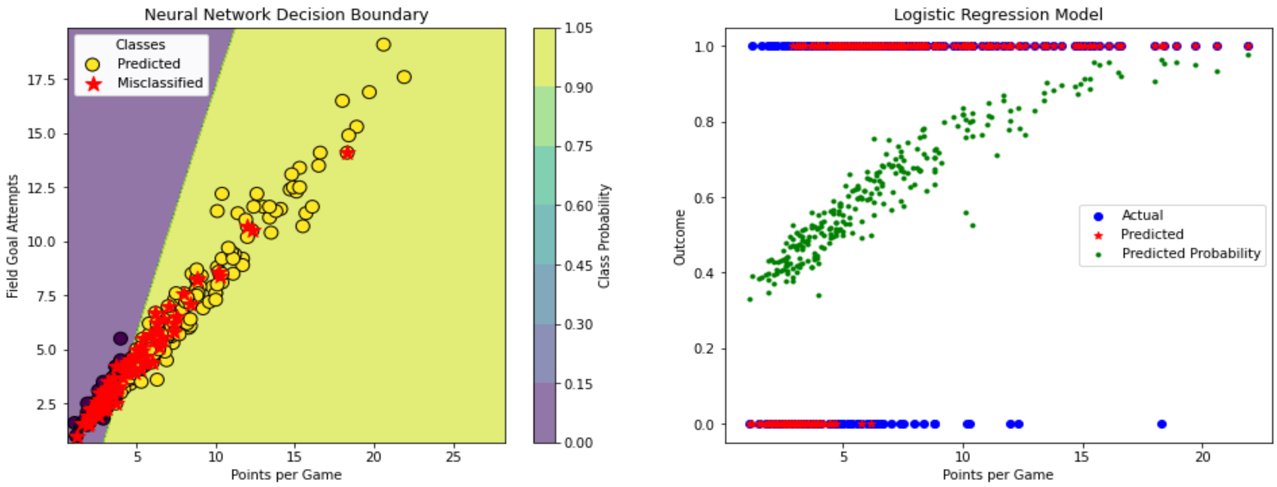

Figure 3.1: Performance of Logistic Regression and ANN

Figure 3.1 illustrates the manner in which classes have been distributed in the Logistic Regression (ANN) model. It identifies the misclassified plots as well as the correctly predicted plots. Class probability has been clearly demarcated which helps us better understand the magnitude of players who lasted 5 years in the NBA and those who did not.

3.3 Complex Models: Data Preprocessing

The entire dataset is utilized for modelling, evaluation, and visualization. After assigning variables, missing values, and outliers are addressed. Histogram analysis reveals right skewness in most features, mitigated by log transformation, achieving a more symmetrical distribution. Standardization ensures each feature has a mean of 0 and a standard deviation of 1, preventing undue influence on the model.

3.4 Model Performance (Complex Models)

Model deployed (All Features)	Accuracy (%)	Mislabelled Points	Total Points
Artificial Neural Network	73.68	70	266
Gaussian Naïve Bayes	68.42	826	1329
Logistic Regression	75	503	1329

Table 3.2: Performance of Models featuring all variables

We notice an increase in the accuracy of all models post-data preprocessing when compared with previous simple models. Accuracy has been approximated to 74% and 68% for ANN and GNB respectively. Interestingly, Logistic Regression has shown the most improvement with 75% accuracy which has been illustrated in Figure 3.2. An AUC (Area Under Curve) value of 0.80 is deemed favorable, indicating that the model is proficient in assigning higher probabilities to randomly selected positive instances compared to randomly chosen negative instances.

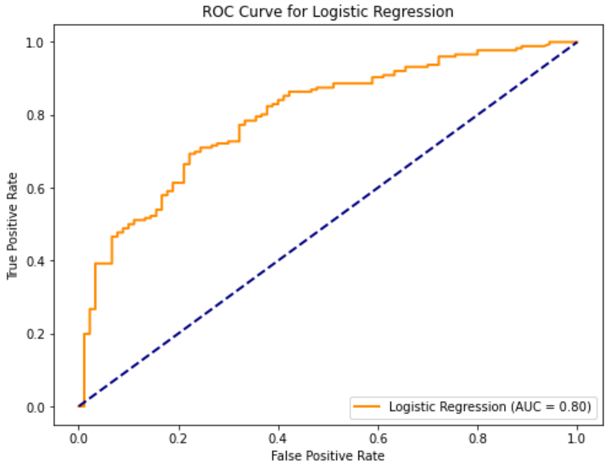

Figure 3.2: Model Performance for Logistic Regression (All Features)

3.5 Suggestions and Further Improvements

It is suggested that the Artificial Neural Network and Logistic Regression be the preferred models of choice. These models can prove to be valuable for predicting player career length. Additionally, it is recommended that a similar analysis be applied at a college league level to categorize player career lengths in the NBA based on their college-level statistics. To ensure fairness, the model should be assessed for bias towards any race, demographic, or nationality.

Additionally, it is recommended that a continuous monitoring cycle be followed with regular updates as the model may change over time. Furthermore, it is suggested that mislabeled points be identified and analyzed for they may contain patterns and trends that can be recognized. This can result in decreasing the magnitude of mislabeled points and further improve the accuracy of the models.

Further improvements to the Logistic Regression and ANN models can be brought by tuning hyperparameters, and experimenting with different architectures and activation functions.

Task 4: Ethics of AI

The Trolley Problem, a time-honored ethical dilemma extends its influence to modern times when it is applied in the context of Autonomous Vehicles. In its root structure, the trolley problem raises ethical questions where a decision has to be made based on two morally conflicting choices. In terms of autonomous vehicles, the trolley problem has evolved into a situation where the algorithm behind the autonomous vehicle has to make split-second decisions whose results could be fatal.

For instance, let us consider an autonomous vehicle which is driving on a narrow road and a a nearby school has its exit on the narrow road. It is the student's leaving time and they are exiting on the narrow road; at the same time, a heavy vehicle enters the narrow road. Both events are identified and recorded by sensors that are equipped on the autonomous vehicle. Therefore, the autonomous vehicle must decide between two scenarios with a high potential of fatality. Firstly, it can save the students or it can save the occupants of the car by avoiding the aforementioned heavy vehicle. The aforementioned instance mimics the ethical complexities faced in the Trolley Problem.

There exists a potential for several key issues to arise from these decisions, which can potentially hinder the adoption of autonomous vehicles. Some of them have been identified as follows:

1) Moral and Societal Issue:

Modern autonomous vehicles are backed by algorithms that are designed to follow rules and protocols. However, they lack emotional intelligence and moral reasoning making it difficult for them to handle morally challenging situations.

Furthermore, it has been opined that there should be universal standards for morally challenging scenarios for autonomous vehicles. However, having a single protocol applied as an umbrella standard globally could result in societal and cultural differences. Additionally, upon the application of area-specific standards and protocol, the question of who decides on ethical parameters be it engineers, policymakers, or the society through a consensus arises (Floridi, 2023).

2) Safety Optimization and a Competitive Environment:

It has been argued by a section of industry leaders and academicians that the algorithm should lead to a situation resulting in the least harm; however, this could compromise the safety of occupants. On the other hand, a self-conservatory approach has been preferred where occupant safety is paramount, potentially resulting in harm to others. Manufacturers can follow a top-down or a bottom-up approach for algorithm development for autonomous vehicles, creating challenges in technological coordination while aiming to remain competitive. Moreover, in the absence of a global standard and protocol, how will manufacturers coordinate with each other in terms of vehicle-to-vehicle communication and interoperability? (Geisslinger *et al.*, 2021)

3) Legal Issue:

Autonomous vehicle manufacturers are likely to face legal issues since a majority of fatal road accidents on a global scale occur in morally challenging situations (Robinson *et al.*, 2021). For instance, under the United States liability law, a victim of a road accident can claim damages by income lost to dependants (Ross, 2023). Thus, in the United States, an autonomous vehicle is expected to drive more carefully in an economically advantaged area compared to a developing area thereby, encouraging biased behavior.

Navigating through the aforementioned issues requires a collaborative approach and efforts where industry, academia, and society converge to find solutions. Technological advancements, legal frameworks, ethical discourse, and consensus are all ingredients required for addressing the issues posed by autonomous vehicles.

References

Çınar, A.C. (2020) 'Training Feed-Forward Multi-Layer Perceptron Artificial Neural Networks with a Tree-Seed Algorithm,' *Arabian Journal for Science and Engineering*, 45(12), pp. 10915–10938. https://doi.org/10.1007/s13369-020-04872-1.

De Souza, L.A. and Costa, H.G. (2022) 'Managing the Conditions for Project Success: An Approach Using k-means Clustering,' in *Springer eBooks*, pp. 396–406. https://doi.org/10.1007/978-3-030-96305-7_37.

Floridi, L., 2023. The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities. The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities - Luciano Floridi - Google Books

Geisslinger, M. *et al.* (2021) 'Autonomous Driving Ethics: from Trolley Problem to Ethics of Risk,' *Philosophy & Technology*, 34(4), pp. 1033–1055. https://doi.org/10.1007/s13347-021-00449-4.

Huang, F.L. (2019) 'Alternatives to logistic regression models in experimental studies,' *Journal of Experimental Education*, 90(1), pp. 213–228. https://doi.org/10.1080/00220973.2019.1699769.

Ismail, S. and Reza, H. (2022) 'Evaluation of naïve Bayesian Algorithms for Cyber-Attacks detection in wireless sensor networks,' *2022 IEEE World AI IoT Congress (AIIoT)* [Preprint]. https://doi.org/10.1109/aiiot54504.2022.9817298.

Maulud, D.H. and Abdulazeez, A.M. (2020) 'A review on Linear Regression comprehensive in Machine Learning,' *Journal of Applied Science and Technology Trends*, 1(4), pp. 140–147. https://doi.org/10.38094/jastt1457.

Robinson, J. *et al.* (2021) 'Ethical considerations and moral implications of autonomous vehicles and unavoidable collisions,' *Theoretical Issues in Ergonomics Science*, 23(4), pp. 435–452. https://doi.org/10.1080/1463922x.2021.1978013.

Ross, D.L., 2023. *Civil liability in criminal justice*. Taylor & Francis. Civil Liability in Criminal Justice - Darrell L. Ross - Google Books