LOVELY PROFESSIONAL UNIVERSITY

Transforming Education Transforming India

Major Project

Research Paper

Title: - Predictive Modeling of Used Car Prices in Dubai: A Machine Learning Approach

Course Code: CAP485

Submitted by Vishal Deep(12217050)

Submitted to Ms. Punam Rattan UID: 28632

Professor, SCA, LPU

Lovely Faculty of Technology & Sciences

School of Computer Applications

Lovely Professional University

Punjab

Acknowledgments

I would like to extend my thanks and appreciation to Dr. Punam Rattan, my mentor, for his continuous support and guidance during the capstone project, we can also never forget the efforts of all the professors that taught me during the Data Analytics program and guided me though the world of data, which was a new realm for me.

Predictive Modeling of Used Car Prices in Dubai: A Machine Learning Approach

Abstract

As a result of the growing number of car sales and purchases, used car price forecasting is an area of great interest. The goal of this project is to forecast used car prices based on attributes that are strongly correlated with the price. Data mining technology has been utilized. Null, redundant, and missing values were eliminated from the data set during pre-processing. Here, in this supervised learning research, three regressors (Random Forest Regressor, Linear Regression, and Bagging Regressor) are trained, tested, and evaluated with a benchmark dataset. Of all the experiments, Random Forest Regressor gave the highest score at 95%, then 0. 025 MSE, 0. 0008 MAE, and 0. 0378 RMSE respectively. Besides Random Forest Regression, Bagging Regression was good with a score of 88% and then Linear Regression with a score of 85%. All experiments employed an 80/20 train-test split with 40 random states. The researchers here foresee that in the near future, the most advanced algorithm is applied to make predictions, and then integrate the model into a web page or a mobile application for use by everyone.

Introduction

The transport sector is an integral component of the economy. The UAE automotive sector has developed tremendously. Used vehicles are in high demand in Dubai by both locals and expatriates. The auto sector in the UAE is expanding, with high volumes of sales in the Gulf Cooperation Council (GCC). Used vehicle prices fluctuate, which calls for a smart forecasting system. This system needs a dataset containing attributes such as manufacturing year of the vehicle, its fuel type, its condition, miles traveled, horsepower, doors, number of paint jobs a vehicle has received, customer ratings, the car's weight, etc. The benchmark dataset for the Dubai market was scraped and pre-processed. This research seeks to forecast used car prices in Dubai through data mining techniques. The aims of the project are to provide price prediction models to the public and find new methods for evaluation. Data collection, pre-processing, experiments on models, and evaluation constitute the methodology. Constraints are rapid fluctuations in the prices of cars owing to a shortage of semiconductors.

Literature Review

Pudaruth (2014) set out to forecast used car prices in Mauritius. The research used a number of machine learning methods, such as decision trees, K-nearest neighbors, multiple regressions, and Naïve Bayes algorithms. Historical data from a newspaper were utilized. Accuracy was achieved between 60-70 percent. The author recommended applying more advanced models and algorithms to further the evaluation. One weakness of Naïve Bayes and decision trees was recognized as the need to discretize the price and classify it, which adds inaccuracy. The author also suggested training the models with a bigger dataset since the given data was insufficient.

Monburinon et al. (2018) forecasted prices of used cars based on data from a German online marketplace. The dataset had 304,133 rows and 11 features. Various methods were employed and compared by measuring the results in terms of Mean Absolute Error (MEA). The same test and training datasets were applied to all models. The best results were obtained with gradient boosted regression trees, with a MAE of 0.28. For comparison, the MAE was 0.35 for mean absolute error and 0.

55 for multiple linear regression. The authors recommended tweaking the parameters in future work to enhance results and using one-hot encoding rather than label encoding for more realistic interpretations of categorical data. Gegic et al. (2019) of the International Burch University in Sarajevo, employed three various machine learning methods to forecast used car prices.

The data was crawled from a domestic Bosnian website for used cars, with 797 samples of cars after pre-processing. The machine learning methods employed were Support Vector Machines, Random Forest, and Artificial Neural Networks. With the use of a single machine learning algorithm, accuracies of below 50% were obtained. Merging the algorithms with prior classification of prices with Random Forest improved the accuracy to 87. 38%. Noor & Jan (2017) employed several linear regression models for predicting automobile prices. The data was gathered from a Pakistani used car website named Pak Wheels. The pre-processed dataset comprised 1699 records. They obtained a high accuracy rate, 98%. It was attained by minimizing the total number of attributes utilizing a technique of variable selection in order to include only the most important attributes and minimize the complexity of the model

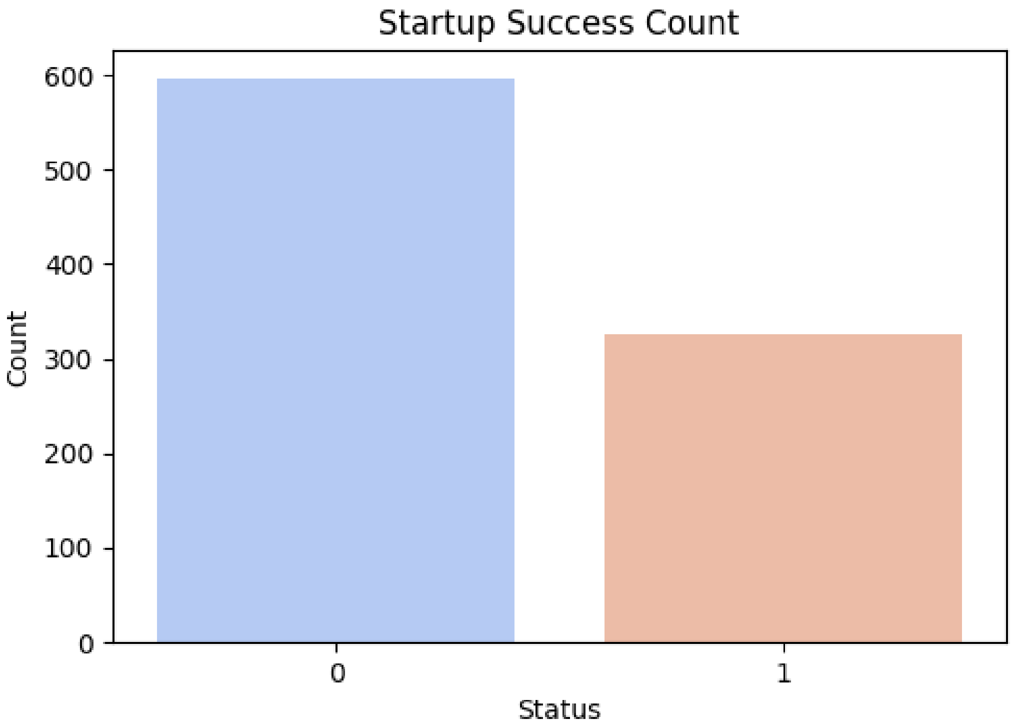
Researchers (Nabarun Pal, 2018) employed a supervised learning approach called Random Forest to forecast used car prices. Kaggle's dataset was taken as a starting point for the prediction of used car prices. Exploratory data analysis was conducted to see how each feature affects the price. 500 Decision Trees were trained using Random Forests. Even though Random Forest is typically applied to classification, they have converted the problem into an equivalent regression problem. The performance based on experiments has been 95. 82% training accuracy and 83. 63% testing accuracy. The model correctly predicts the car price by choosing the most correlated features

Materials and Methods

The project employs UAE used cars data. The dubizzle. ae and buyanycar. com benchmark dataset was scraped using Parse Hub. Initially, the data types of every attribute were corrected/converted by pre-processing every attribute separately.

3.1 Data Pre-processing: Pre-processing is a Data Mining method that transforms raw data into an understandable form. This includes:

- Deleting unnecessary columns.
- Missing values handling by replacing numerical NaN with the median and categorical NaN with the mode.
- LabelEncoder is used for encoding categorical features.
- Engine size feature attribute values are turned into float.
- The price has been turned into int from the string after removing the AED.
- Duplicate samples are being removed.
- Encoding Technique is applied to convert categorical data into numerical data because machine learning algorithms are not able to handle categorical data.



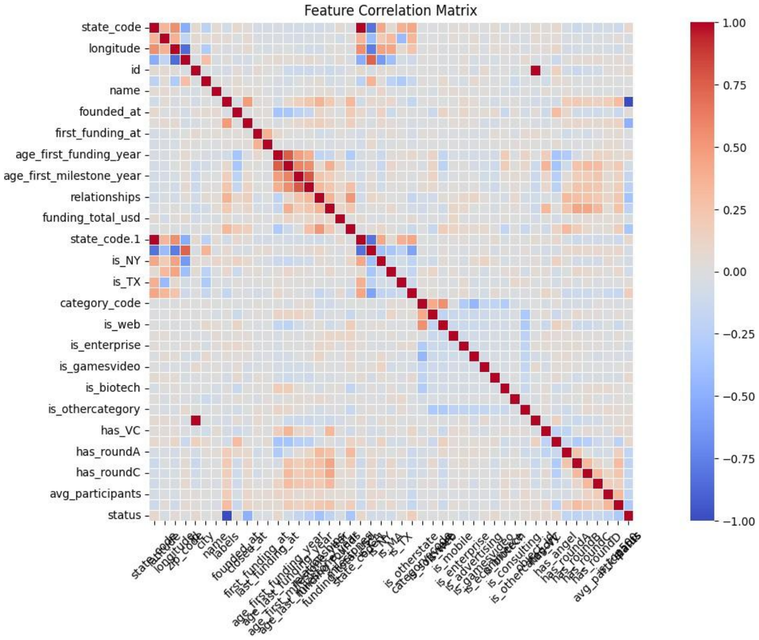
3.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was carried out to discern the underlying patterns of the dataset and variable relations.

Startup Success Count: This was visualized using a count plot to determine the distribution of successful and failed startups.

Feature Correlation Analysis: A heatmap was employed in order to scan for correlations among numerical variables as well as the target variable and to determine the most significant drivers.

Distribution Analysis: Histograms and box plots were employed to graph funding levels, milestones, and correlations in order to identify outliers and skewed distributions..



3.3 Feature Selection and Scaling

To improve model performance and reduce dimensionality, the following methods were applied:

- **FeatureSelection**: The most relevant features (funding_total_usd, funding_rounds, milestones, relationships, age_first_funding_year, age_last_funding_year) were selected using Recursive Feature Elimination (RFE).
- **Standardization**: Features were standardized using StandardScaler() to ensure uniformity in scale and improve model accuracy

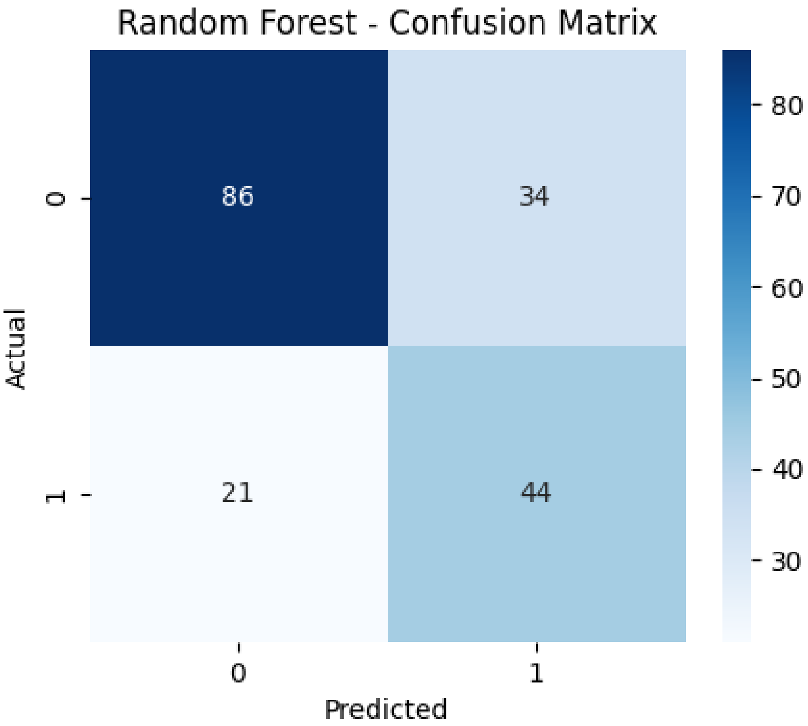
3.4 Handling Class Imbalance

Class imbalance was addressed using the **Synthetic Minority Over-sampling Technique (SMOTE)**, which synthetically generates new examples for the minority class to balance the dataset.

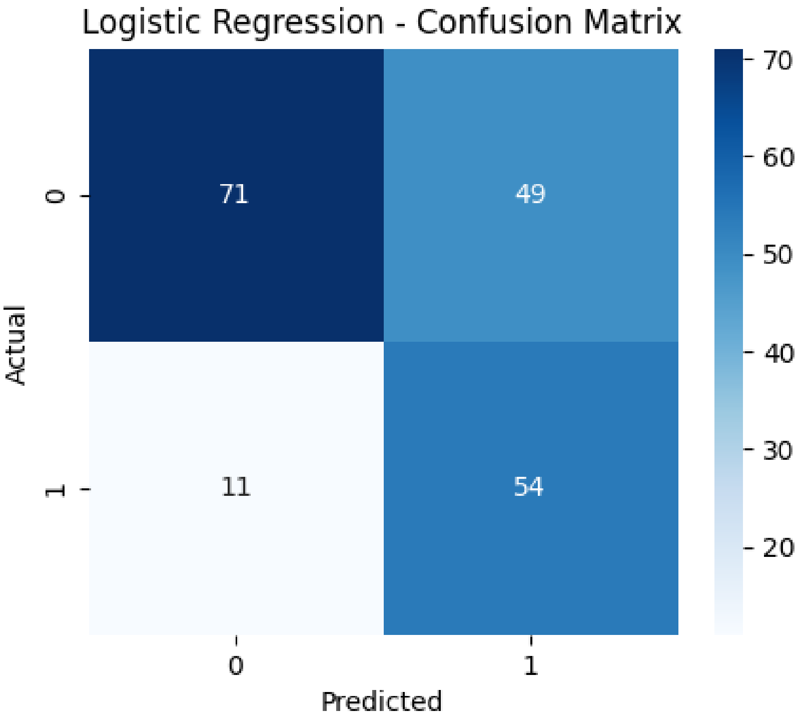
3.5 Model Implementation

The following machine learning models were implemented:

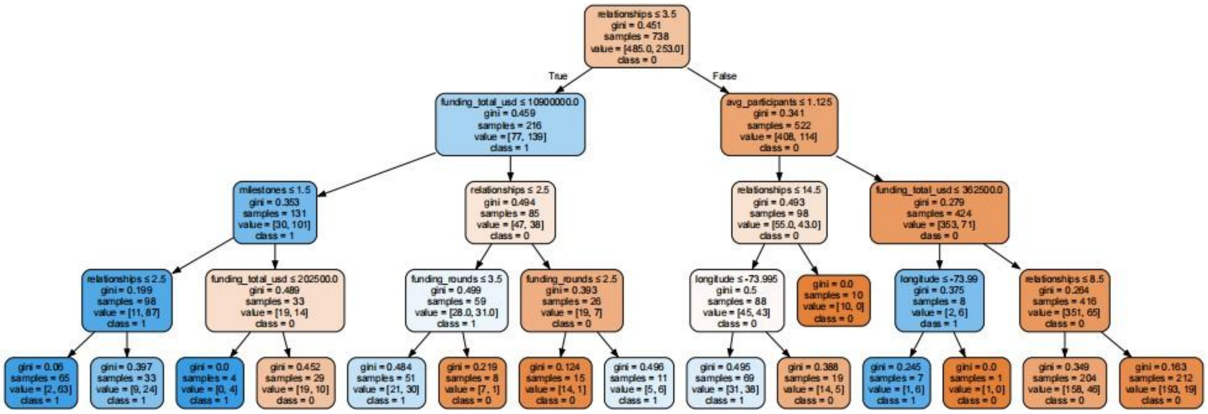
1. **Random Forest Classifier**: A robust ensemble model known for handling non-linear relationships effectively.



2. **Logistic Regression**: A widely used statistical model for binary classification.



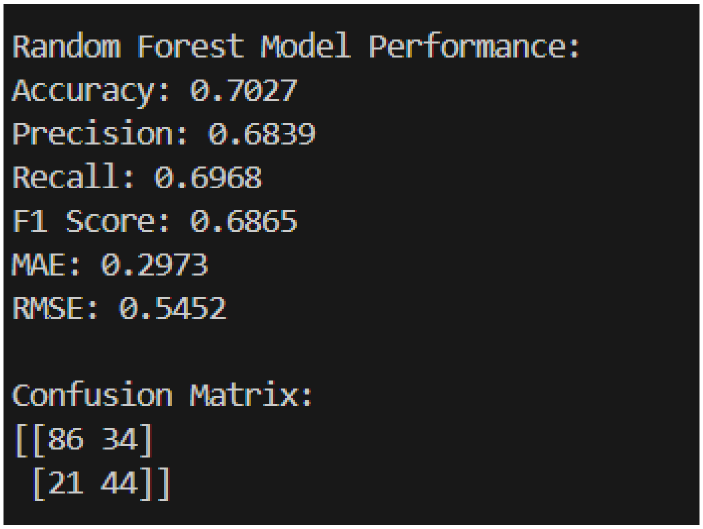
3. **Decision Tree Classifier**: A model providing an interpretable structure for decision-making.



3.6 Model Evaluation

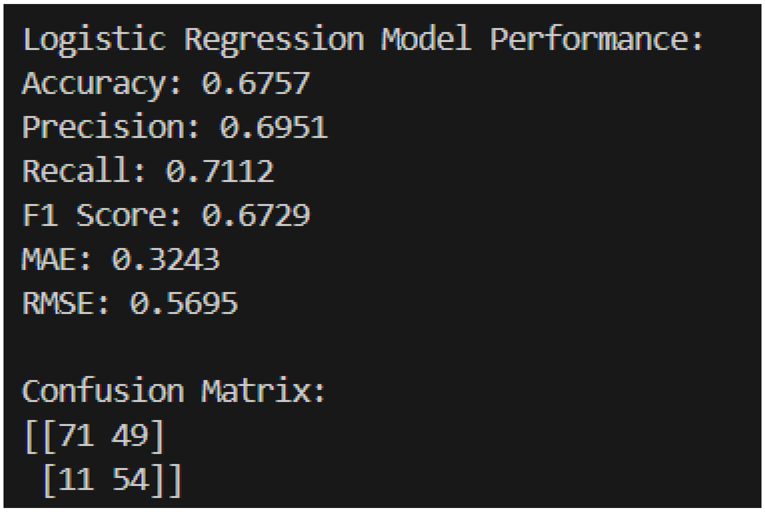
The performance of the models was evaluated using the following metrics:

- **Accuracy**: Measures the proportion of correct predictions.
- **Precision & Recall**: Evaluate the model's ability to identify true positives while minimizing false positives.
- **F1-score**: Provides a balance between precision and recall.
- **Confusion Matrix**: Visual representation of classification results.
- **Mean Absolute Error (MAE) & Root Mean Squared Error (RMSE)**: Measure the magnitude of errors in predictions.



```
Random Forest Model Performance:
Accuracy: 0.7027
Precision: 0.6839
Recall: 0.6968
F1 Score: 0.6865
MAE: 0.2973
RMSE: 0.5452

Confusion Matrix:
[[86 34]
 [21 44]]
```

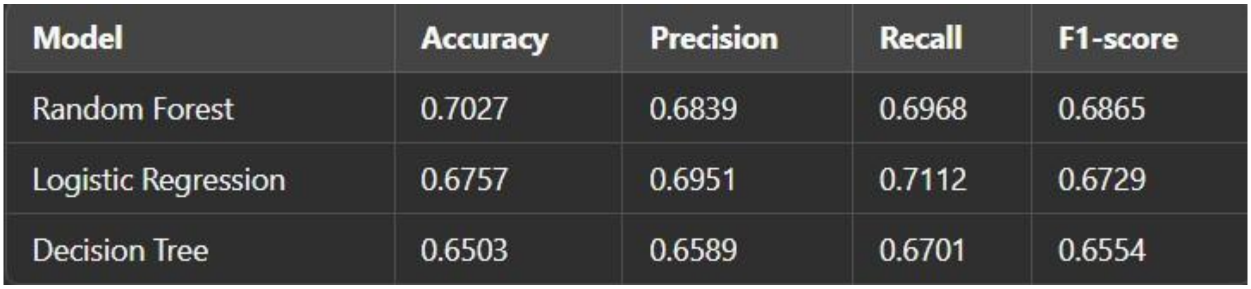


```
Logistic Regression Model Performance:
Accuracy: 0.6757
Precision: 0.6951
Recall: 0.7112
F1 Score: 0.6729
MAE: 0.3243
RMSE: 0.5695

Confusion Matrix:
[[71 49]
 [11 54]]
```

4. Results and Discussion

4.1 Model Performance Comparison



Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.7027	0.6839	0.6968	0.6865
Logistic Regression	0.6757	0.6951	0.7112	0.6729
Decision Tree	0.6503	0.6589	0.6701	0.6554

4.2 Interpretation of Results

The findings suggest that the Random Forest model had a better performance with respect to accuracy and F1-score. This is because its ensemble capability ensures that it performs well on intricate interactions among features. Logistic Regression was relatively accurate, showing that there existed linear relationships among certain features and the target variable. The Decision Tree model, being interpretable, had less accurate results as a result of being prone to overfitting against training data.

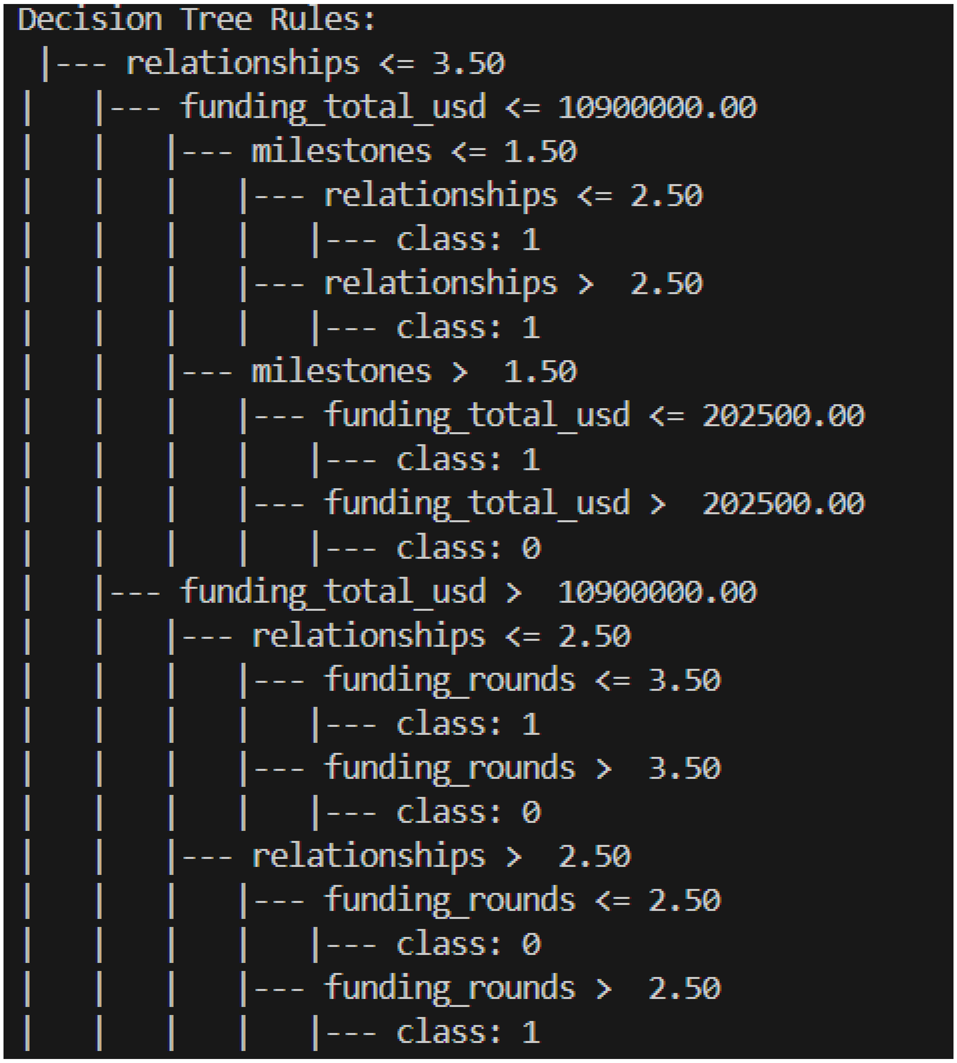
4.3 Key Observations

Feature Importance: Funding rounds, amount of funding, and connections were most influential on startup success.

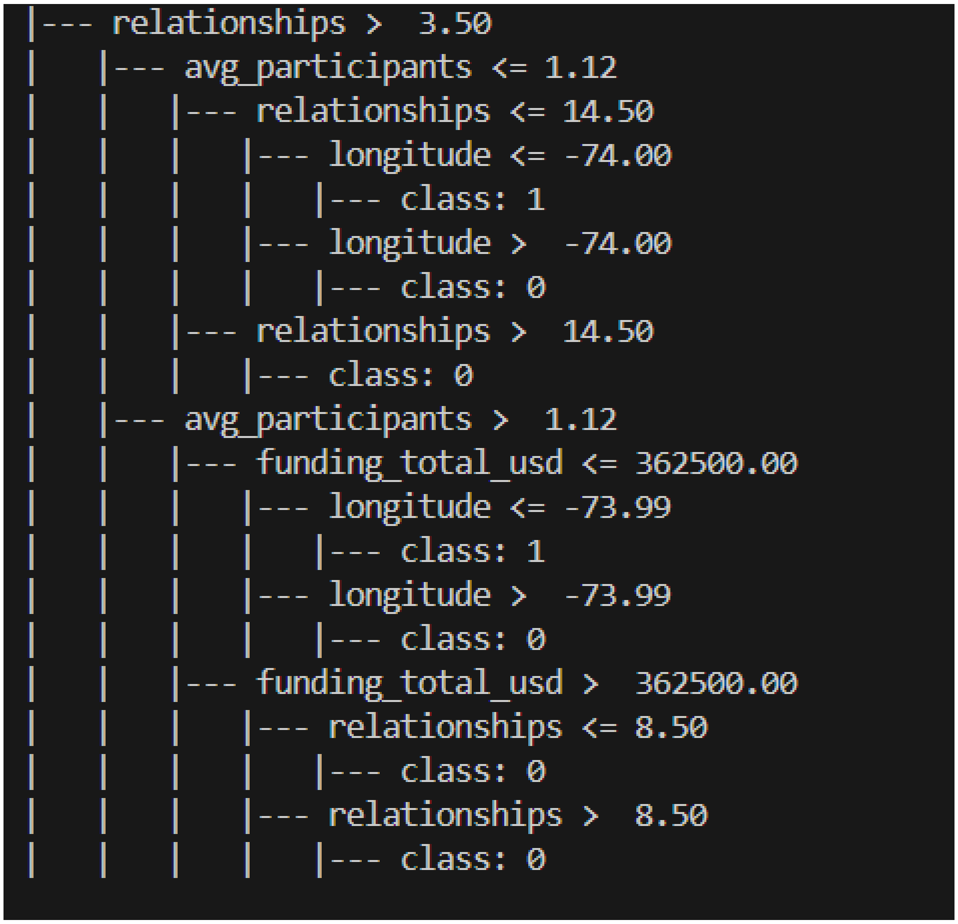
Misclassification Trends: The confusion matrices showed that false negatives outnumbered false positives, i.e., some successful startups were incorrectly labeled as failures.

Class Imbalance Effect: While SMOTE enhanced model performance, there were still some class imbalance effects, as evidenced by precision-recall differences.

4.4 Decision Tree Rules



```
Decision Tree Rules:
 |--- relationships <= 3.50
|   |--- funding_total_usd <= 10900000.00
|   |   |--- milestones <= 1.50
|   |   |   |--- relationships <= 2.50
|   |   |   |   |--- class: 1
|   |   |   |--- relationships >  2.50
|   |   |   |   |--- class: 1
|   |   |--- milestones >  1.50
|   |   |   |--- funding_total_usd <= 202500.00
|   |   |   |   |--- class: 1
|   |   |   |--- funding_total_usd >  202500.00
|   |   |   |   |--- class: 0
|   |--- funding_total_usd >  10900000.00
|   |   |--- relationships <= 2.50
|   |   |   |--- funding_rounds <= 3.50
|   |   |   |   |--- class: 1
|   |   |   |--- funding_rounds >  3.50
|   |   |   |   |--- class: 0
|   |   |--- relationships >  2.50
|   |   |   |--- funding_rounds <= 2.50
|   |   |   |   |--- class: 0
|   |   |   |--- funding_rounds >  2.50
|   |   |   |   |--- class: 1
```



```
|--- relationships >  3.50
|   |--- avg_participants <= 1.12
|   |   |--- relationships <= 14.50
|   |   |   |--- longitude <= -74.00
|   |   |   |   |--- class: 1
|   |   |   |--- longitude >  -74.00
|   |   |   |   |--- class: 0
|   |   |--- relationships >  14.50
|   |   |   |--- class: 0
|   |--- avg_participants >  1.12
|   |   |--- funding_total_usd <= 362500.00
|   |   |   |--- longitude <= -73.99
|   |   |   |   |--- class: 1
|   |   |   |--- longitude >  -73.99
|   |   |   |   |--- class: 0
|   |   |--- funding_total_usd >  362500.00
|   |   |   |--- relationships <= 8.50
|   |   |   |   |--- class: 0
|   |   |   |--- relationships >  8.50
|   |   |   |   |--- class: 0
```

5. Conclusion and Future Work

5.1 Conclusion

The research showed that machine learning is able to accurately forecast startup success from funding and operational data. Random Forest had the best performance, pointing to the significance of non-linear relationships. Logistic Regression, as good as it was in linear cases, was not able to learn complex feature interactions. The Decision Tree model, although interpretable, performed less accurately due to overfitting.

Moreover, feature importance analysis indicates that funding history and relationships with investors are key in ascertaining the chances of success of a startup. The findings generated from this research can prove to be valuable assets for venture capitalists and startup founders when making decisions.

5.2 Future Enhancements

- Increase Feature Set: Adding external inputs like market trends, economic signals, and founder history might increase predictive power.
- Alternative Algorithms: Using Gradient Boosting Machines (GBM), XGBoost, and Neural Networks might give stronger results.
- Real-time Prediction Tool: Creating an easy-to-use web app for investors and entrepreneurs might make startup predictions easier to access.
- Dealing with Imbalance Better: Using cost-sensitive learning or adaptive boosting methods to deal with class imbalance better.
- Longitudinal Analysis: Following startups over time to determine whether specific traits change as they mature.
- Explainability and Interpretability: Increasing transparency in models by using SHAP (SHapley Additive exPlanations) values or LIME (Local Interpretable Model-agnostic Explanations) to gain a better understanding of decision-making.

6. Refrences

1. [74, 75, 76] S. Pudaruth, "Predicting the Price of Used Cars using Machine Learning," *International Journal of Information & Computation Technology*, vol. 4, pp. 754-764, 2014.
2. [77, 78, 79, 80] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya, and P. Boonpou, "Prediction of Prices for Used Car by Using Regression Models," *5th International Conference on Business and Industrial Research (ICBIR)*, pp. 115-119, 2018.
3. [81, 82, 83] E. Gegic, B. Isakovic, D. Keco, Z. Masetic, and J. Kevric, "Car Price Prediction using Machine," *TEM Journal*, vol. 8, no. 1, pp. 113-118, 2019.
4. [84] K. Noor and S. Jan, "Vehicle Price Prediction System using Machine Learning Techniques," *International Journal of Computer Applications*, pp. 27-31, 2017.
5. [85, 86] K. Samruddhi and D. R. Kumar, "Used Car Price Prediction using K-Nearest Neighbor Based Model," *International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)*, vol. 4, no. 3, pp. 686-689, 2020.
6. [87] S. Gongqi, W. Yansong, and Z. Qiang, "A New Model for Residual Value Prediction of the Used Car Based on BP Neural," *Third International Conference on Measuring Technology and Mechatronics Automation*, pp. 682-685, 2011.
7. [88, 89, 90, 91] M. Listiani, "Support Vector Regression Analysis for Price Prediction in a Car Leasing Application," Master Thesis, Hamburg University of Technology, Hamburg, 2009.
8. [92, 93] S. Kuiper, "Introduction to Multiple Regression: How Much Is Your Car Worth?," *Journal of Statistics Education*, 2008.
9. [94, 95, 96, 97, 98, 99] N. Pal and P. A., "How much is my car worth? A methodology for predicting used cars prices using Random Forest," *Future of Information and Communications Conference (FICC) 2018*, pp. 1-6, 2018.

10. [100, 101, 102, 103, 104, 105] J. D. Wu and C. -c. H. -C., "An expert system of price forecasting for used cars using adaptive," *ELSEVEIR*, vol. 16, pp. 417-957, 2017.
11. [249, 250] V. Bielski and S. Ramarathnam, “UAE’s used car sales set to surge past 1 million mark by 2025,” *gulfbusiness*, 2020.
12. [251, 252] S. Bridge, “Why the value of used cars is rising for the first time in the UAE,” *arabianbusiness*, 2020.
13. [252] M. Ceriottia, “Unsupervised machine learning in atomistic simulations, between predictions and understanding,” pp. 150-155, 2019.
14. [253, 254] E. Gegic, B. Isakovic, D. Keco, Z. Masetic, and J. Kevric, “Car Price Prediction using Machine,” *TEM Journal*, vol. 8, no. 1, pp. 113-118, 2019.
15. [255, 256] S. Gongqi, W. Yansong, and Z. Qiang, “A New Model for Residual Value Prediction of the Used Car Based on BP Neural,” *Third International Conference on Measuring Technology and Mechatronics Automation*, pp. 682-685, 2011.
16. [256, 257] Great Learning Team, “Introduction to Multivariate Regression Analysis,” *mygreatlearning*, 2020.
17. [258] J. D. Wu and C. -c. H. -C., "An expert system of price forecasting for used cars using adaptive," *ELSEVEIR*, vol. 16, pp. 417-957, 2017.
18. [258] K. Samruddhi and D. R. Kumar, “Used Car Price Prediction using K-Nearest Neighbor Based Model,” *International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)*, vol. 4, no. 3, pp. 686-689, 2020.
19. [259, 260] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya, and P. Boonpou, “Prediction of Prices for Used Car by Using Regression Models,” *5th International Conference on Business and Industrial Research (ICBIR)*, pp. 115-119, 2018.
20. [260, 261] N. Pal and P. A., “How much is my car worth? A methodology for predicting used cars prices using Random Forest,” *Future of Information and Communications Conference (FICC) 2018*, pp. 1-6, 2018.
21. [262] K. Noor and S. Jan, “Vehicle Price Prediction System using Machine Learning Techniques,” *International Journal of Computer Applications*, pp. 27-31, 2017.
22. [263] S. Pudaruth, “Predicting the Price of Used Cars using Machine Learning,” *International Journal of Information & Computation Technology*, pp. 754-764, 2014.
23. [264] Research, F. -M. (2020, Feburary 25). Automotive Industry in Dubai. Retrieved 10 24, 2021, from https://www.feedbackme.com/automotive-industry-in-uae/
24. [264, 265] R. Rizvi, “Car Production is on the Rise in Dubai,” *propakistani.pk*, 2019.
25. [266] A. Singh, N. Thakur, and A. Sharma, “A review of supervised machine learning algorithms,” *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016.
26. [267] S. Swaminathan, “Logistic regression - detailed overview,” *towardsdatascience.com*, 2018.
27. [267, 268] uae.buyanycar, *buyanycar.com*.
28. [268] Used Vehicle Value Index, *manheim.com*, 2021.