- Built and deployed a classification model to predict whether a passenger will refer the airline to others, addressing a critical business need for improving customer satisfaction and retention.
- Handled imbalanced classes through oversampling techniques, tackled missing values, and implemented feature engineering to improve prediction accuracy.
- Achieved a model accuracy of 87% using Random Forest and a precision score of 85% for the positive class.
- Deployed the model using Streamlit, allowing for interactive testing and feedback from business users, with a potential to increase referral rates by 20% based on prediction analysis.

Keywords in CSV format:
Classification, Random Forest, Decision Tree, Data Preprocessing, Feature Engineering, Class Imbalance, Oversampling, Accuracy, Precision, Model Evaluation, Streamlit Deployment, Customer Retention, Predictive Analytics, Scikit-learn, Python, Cross-Validation, Model Optimization, Business Impact, User Interaction, Referral Prediction

10 Project-Specific Questions:

1. **What was the main objective of the Airline Referral Prediction project?**
   - The objective was to predict whether a passenger referred by an existing customer would book a flight based on features like seat comfort, cabin service, travel class, food & beverage, and entertainment service.
2. **How did you preprocess the data, and what transformations were necessary?**
   - Preprocessing involved handling missing values, converting data types for model compatibility, and normalizing or encoding categorical variables.
3. **What were the main challenges faced during data preprocessing?**
   - Challenges included handling missing values in categorical variables, standardizing ratings, and dealing with imbalanced class distribution.
4. **Which features had the most significant impact on the prediction, and why?**
   - Features like Overall Rating, Value for Money, and Recommendations were most impactful, as they directly reflect customer satisfaction and likelihood of referral.
5. **What insights did you derive from the Exploratory Data Analysis (EDA)?**
   - Key insights included identifying a high percentage of solo travelers, low satisfaction with entertainment services, and a significant number of overall ratings below 3.0.
6. **What was the reason for choosing classification models for this problem?**
   - The problem was a binary classification task, predicting whether a passenger would refer the airline, making classification models a suitable choice.

7. **Which machine learning models did you test, and what were their results?**
   - Models such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbour, Support Vector Machine, and Naive Bayes were tested. Logistic Regression had the best balance of performance and interpretability.
8. **What was your approach to hyperparameter tuning, and how did it affect model performance?**
   - Grid Search CV was used for hyperparameter tuning, which optimized model parameters and helped in achieving a higher accuracy.
9. **What evaluation metrics did you use to measure model performance?**
   - Metrics like Accuracy, Precision, Recall, F1-Score, and ROC-AUC were used for performance measurement.
10. **What were your recommendations to the airline based on the model's results?**
    - Recommendations included focusing on enhancing cabin service, ground service, food & beverage offerings, and seat comfort to improve overall ratings and referral rates.

---

40 Data Science Technical Interview Questions and Answers:

1. What is the difference between supervised and unsupervised learning?

- **Supervised Learning:** Uses labeled data to train models (e.g., classification, regression).
- **Unsupervised Learning:** Uses unlabeled data to find hidden patterns (e.g., clustering, association).

2. Explain overfitting and underfitting. How can you mitigate them?

- **Overfitting:** When the model performs well on training data but poorly on new data. Mitigated using regularization, reducing model complexity, or using more data.
- **Underfitting:** When the model is too simple to capture patterns. Solved by increasing model complexity or using more relevant features.

3. What is cross-validation? Why is it used?

- Cross-validation splits the data into multiple folds to train and validate the model multiple times. It prevents overfitting and ensures model generalization.

4. What are Precision and Recall? How are they calculated?

- **Precision:** TP/(TP+FP)\text{TP} / (\text{TP} + \text{FP})TP/(TP+FP)
- **Recall:** TP/(TP+FN)\text{TP} / (\text{TP} + \text{FN})TP/(TP+FN)
- Where TP = True Positive, FP = False Positive, and FN = False Negative.

5. Explain the bias-variance tradeoff.

- **Bias:** Error due to overly simplistic assumptions in the learning algorithm.

- **Variance:** Error due to too much complexity in the learning algorithm.
- Tradeoff: Increasing bias decreases variance and vice-versa. The goal is to find the optimal balance.

6. What is regularization? Explain L1 and L2 regularization.

- Regularization prevents overfitting by adding a penalty to the loss function:
- **L1 (Lasso):** Adds absolute value of coefficients. Promotes sparsity.
- **L2 (Ridge):** Adds square of coefficients. Encourages small weights.

7. Explain Principal Component Analysis (PCA).

- PCA reduces the dimensionality of data by transforming features into a set of linearly uncorrelated components, maximizing variance along each component.

8. What is feature engineering?

- Feature engineering involves creating new features or modifying existing ones to improve model performance. Techniques include normalization, encoding, and feature interaction.

9. How do you handle imbalanced datasets?

- Techniques include:
  - **Resampling:** Oversampling the minority class or undersampling the majority class.
  - **SMOTE (Synthetic Minority Over-sampling Technique).**
  - **Using appropriate evaluation metrics** (Precision, Recall, F1-Score).

10. What is the ROC-AUC score?

- ROC-AUC score measures a model's ability to distinguish between classes. It plots True Positive Rate vs. False Positive Rate, with a value closer to 1 indicating a better model.

11. Explain k-fold cross-validation.

- Splits data into k parts, trains on k-1 parts, and tests on the remaining part. This process is repeated k times to get an average performance measure.

12. What is gradient descent?

- Gradient Descent is an optimization algorithm used to minimize the loss function by iteratively updating model parameters in the opposite direction of the gradient.

13. Explain the concept of Ensemble Learning.

- Combines multiple models to produce a better result. Techniques include:
  - **Bagging (e.g., Random Forest).**
  - **Boosting (e.g., AdaBoost, XGBoost).**

14. How do Decision Trees work?

- Decision Trees recursively split the data based on feature values, choosing splits that maximize information gain or minimize Gini impurity.

15. What are support vector machines (SVM)?

- SVMs classify data by finding a hyperplane that maximizes the margin between different classes. Support vectors are data points closest to the hyperplane.

16. What is clustering? Name a few clustering algorithms.

- Clustering groups similar data points. Algorithms include K-means, DBSCAN, and Hierarchical Clustering.

17. What is a confusion matrix?

- A confusion matrix summarizes classification results:
  - **TP:** Correctly predicted positive.
  - **TN:** Correctly predicted negative.
  - **FP:** Incorrectly predicted positive.
  - **FN:** Incorrectly predicted negative.

18. Explain Naive Bayes classifier.

- Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming independence among features. It’s effective for text classification tasks.

19. What is the difference between bagging and boosting?

- **Bagging:** Reduces variance by training multiple models in parallel.
- **Boosting:** Reduces bias by training models sequentially, each focusing on errors of the previous one.

20. What is a Random Forest?

- An ensemble of Decision Trees, where each tree is trained on a random subset of features. The final prediction is made by majority voting (classification) or averaging (regression).

20 More Data Science Technical Interview Questions and Answers

21. What is the purpose of feature scaling, and what techniques are used?

- **Purpose:** Feature scaling standardizes or normalizes data so that features contribute equally to the model.
- **Techniques:**
  - **Standardization:** Rescales to have mean = 0 and standard deviation = 1.
  - **Normalization:** Scales features between 0 and 1.

22. Explain the difference between fit(), transform(), and fit_transform() in scikit-learn.

- **fit():** Calculates parameters (e.g., mean, variance) for the transformation.

- **transform():** Applies the transformation to the data.
- **fit_transform():** Combines fit() and transform() in a single step.

23. What is a confusion matrix, and how do you interpret it?

- A confusion matrix is a table that describes the performance of a classification model:
  - **True Positive (TP)**: Correctly predicted positives.
  - **True Negative (TN)**: Correctly predicted negatives.
  - **False Positive (FP)**: Incorrectly predicted positives (Type I Error).
  - **False Negative (FN)**: Incorrectly predicted negatives (Type II Error).

24. What is the curse of dimensionality?

- As the number of features increases, the volume of the feature space grows exponentially, making it difficult to visualize, analyze, and build models. It can cause overfitting and increased computational costs.

25. How does K-Means clustering work?

- **Steps:**
  1. Initialize k centroids randomly.
  2. Assign each data point to the nearest centroid.
  3. Recalculate the centroid of each cluster.
  4. Repeat until convergence (no changes in centroids).

26. What is an Activation Function in neural networks?

- An activation function introduces non-linearity to the network, allowing it to learn complex patterns. Common ones include:
  - **Sigmoid:** Maps input between 0 and 1.
  - **ReLU:** Sets negative values to zero.
  - **Tanh:** Maps input between -1 and 1.

27. What is a convolutional neural network (CNN)?

- A CNN is a deep learning model primarily used for image recognition tasks. It consists of convolutional layers, pooling layers, and fully connected layers that help identify spatial hierarchies in images.

28. What is a Recurrent Neural Network (RNN)?

- An RNN is a neural network that handles sequential data using its internal memory to process variable-length input sequences, making it suitable for time series, text, and speech data.

29. Explain Gradient Boosting.

- Gradient Boosting builds models sequentially, where each new model corrects errors made by previous ones by optimizing a loss function. Popular implementations include XGBoost and LightGBM.

30. What are Autoencoders?

- Autoencoders are neural networks used for unsupervised learning. They learn to compress input data into a latent space representation and then reconstruct the data, often used for anomaly detection or denoising.

31. What is a Markov Chain?

- A Markov Chain is a stochastic model describing a sequence of events, where the probability of each event depends only on the state attained in the previous event (memoryless property).

32. How do you handle multicollinearity in regression models?

- Techniques include:
  - **Dropping highly correlated features.**
  - **Using dimensionality reduction (PCA).**
  - **Applying regularization techniques like Ridge regression.**

33. What is the difference between Gini Impurity and Entropy in Decision Trees?

- Both measure the homogeneity of the nodes:
  - **Gini Impurity:** Probability of misclassifying a random sample. Ranges from 0 (pure) to 0.5.
  - **Entropy:** Measures information gain. Ranges from 0 (pure) to 1.

34. Explain the difference between Bagging and Stacking.

- **Bagging:** Averages multiple independent models to reduce variance.
- **Stacking:** Combines multiple models by training a meta-learner on their predictions to improve performance.

35. What is Time Series Analysis?

- Time Series Analysis involves analyzing data points collected over time. Techniques include ARIMA, Exponential Smoothing, and LSTM networks.

36. How do you choose the right number of clusters in K-Means?

- Using the **Elbow Method** or **Silhouette Score**:
  - **Elbow Method:** Plot WCSS (Within-Cluster Sum of Squares) vs. number of clusters and find the "elbow" point.
  - **Silhouette Score:** Measures the similarity of data points within clusters. Higher scores indicate better clustering.

37. What is the difference between Batch Gradient Descent and Stochastic Gradient Descent?

- **Batch Gradient Descent:** Uses the entire dataset to compute gradients, making it stable but slower.
- **Stochastic Gradient Descent (SGD):** Uses one data point at a time, making it faster but noisier.

38. What are the differences between Type I and Type II errors?

- **Type I Error (False Positive):** Rejecting a true null hypothesis.
- **Type II Error (False Negative):** Failing to reject a false null hypothesis.

39. What is a P-value?

- A P-value indicates the probability of obtaining test results at least as extreme as the results observed, under the assumption that the null hypothesis is true. A lower P-value (< 0.05) suggests strong evidence against the null hypothesis.

40. What is Ridge and Lasso Regression?

- **Ridge Regression (L2 Regularization):** Adds a penalty equal to the square of the magnitude of coefficients.
- **Lasso Regression (L1 Regularization):** Adds a penalty equal to the absolute value of coefficients, promoting sparsity.