|  |  |
| --- | --- |
| Internship Project Title | TCS ION RIO 125: RANK FEATURE OF SMARTPHONE |
| Name of the Company | TCS iON |
| Name of the Industry Mentor |  |
| Name of the Institute | Sanskriti university |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
| 7-09-2024 | 10-10-2024 | 125 | Chrome, Windows 11, ANACONDA ENVIROMENT | MS Office, JUPYTER NOTEBOOK |

**TABLE OF CONTENT**

**Acknowledgement**

I am highly indebted to TCS iON for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards my parents and my academic mentor, for their kind co-operation and encouragement which help me in completion of this project.

I would like to express my special gratitude and thanks to my industry mentor for giving me such attention and time.

**Objective**

The objective of this project is to analyze the features of smartphones and rank them based on their contribution to the pricing and consumer preference. By identifying key features such as battery life, RAM, camera quality, and connectivity options, we aim to create a ranking system that provides insights into consumer priorities when selecting smartphones. This analysis will leverage data-driven methodologies, including logistic regression and visualization techniques, to present findings that are beneficial for consumers and manufacturers alike..

**Introduction / Description of Internship**

During my internship at TCS iON, I had the opportunity to work on an exciting project titled "TCS iON RIO 125 – Rank Features of a Smartphone." This project aimed to analyze the various features of smartphones and rank them based on their significance to pricing and user preference. In an era where technology is evolving rapidly, understanding the attributes that consumers value in smartphones is crucial for both manufacturers and buyers.

Throughout the internship, I was actively involved in multiple facets of data analysis, including data collection, preprocessing, and the application of statistical techniques. My responsibilities included:

**Data Collection** and Cleaning: I gathered data from diverse sources, including smartphone specifications and user reviews. This involved cleaning and preprocessing the data to ensure its quality and relevance for analysis.

**Feature Analysis**: I analyzed various smartphone features such as battery power, RAM, camera specifications, connectivity options (4G, Bluetooth), and additional functionalities (like dual SIM capability and touch screen). This analysis helped identify which features were most influential in determining smartphone prices.

**Ranking Features**: Using techniques like logistic regression, I ranked the smartphone features based on their impact on pricing and user preference. This involved creating a model that could predict price ranges based on specific features, enabling a clearer understanding of consumer priorities.

**Data Visualization:** To effectively communicate my findings, I employed various data visualization tools such as Plotly and Seaborn. This allowed me to create compelling visual representations of the data, highlighting trends and correlations in a user-friendly manner.

**Collaborative Learning:** I collaborated closely with a team of data analysts and mentors, who provided valuable insights and guidance throughout the project. This experience enhanced my skills in data analytics, problem-solving, and teamwork.

**Conclusion**

Overall, my internship at TCS iON provided me with a comprehensive understanding of data analysis in the context of the smartphone industry. It equipped me with practical skills and knowledge that are essential for a successful career in data analytics. I am excited to apply what I learned during this internship to future projects and endeavors in the field of technology and analytics.

**Internship Activities**

**1. \*\*Data Collection and Preparation\*\***

**- \*\*Sourcing Data\*\*:** Collected data from multiple sources, including online databases, industry reports, and market research platforms, focusing on smartphone specifications and user reviews.

**- \*\*Data Cleaning\*\*:** Utilized Python libraries (such as Pandas) to clean the dataset by removing duplicates, handling missing values, and normalizing data formats to ensure accuracy and consistency.

1. **\*\*Feature Analysis\*\***

**- \*\*Identifying Key Features\*\*:** Conducted exploratory data analysis (EDA) to identify key smartphone features, such as battery life, camera quality, RAM, and connectivity options (e.g., 4G, Bluetooth).

**- \*\*Statistical Analysis\***\*: Employed statistical techniques to evaluate the significance of different smartphone features in influencing market prices.

**3. \*\*Ranking Features\*\***

**- \*\*Model Development\*\*:** Developed a logistic regression model to predict smartphone price ranges based on identified features, allowing for a systematic ranking of features by their importance.

**- \*\*Feature Ranking\*\*:** Generated insights on how specific features correlated with pricing, leading to the development of a ranked list of smartphone features.

**4. \*\*Data Visualization\*\***

**- \*\*Creating Visualizations\*\*:** Utilized tools like Plotly and Seaborn to create interactive and static visualizations, including pie charts and distribution plots, to effectively communicate findings to stakeholders.

**- \*\*Presentation of Results\*\*:** Prepared visual presentations of the ranked features and their impacts on pricing, highlighting trends and consumer preferences in the smartphone market.

**5. \*\*Collaborative Learning and Meetings\*\***

**- \*\*Team Collaboration\*\*:** Engaged in regular meetings with team members to discuss findings, share insights, and receive feedback on analyses and methodologies.

**- \*\*Knowledge Sharing\*\*:** Participated in knowledge-sharing sessions led by mentors, gaining insights into advanced analytics techniques and best practices in the industry.

**6. \*\*Documentation and Reporting\*\***

**- \*\*Writing Reports\*\*:** Documented the entire analysis process, including methodologies, findings, and recommendations, in a comprehensive report to be shared with the management team.

**- \*\*Final Presentation\*\*:** Presented the results of the project to stakeholders, detailing the significance of ranked features and their potential implications for future smartphone designs and marketing strategies.

**7. \*\*Skill Development\*\***

**- \*\*Technical Skills\*\*:** Improved proficiency in data analytics tools and programming languages, including Python and its libraries (Pandas, NumPy, Scikit-learn).

**- \*\*Soft Skills\*\*:** Enhanced communication and presentation skills through regular interactions with team members and stakeholders**.**

**Approach / Methodology**

**1. \*\*Problem Definition\*\***

- Clearly defined the objective of the project, which was to rank smartphone features based on their impact on price and user preferences.

- Identified key questions to address, such as which features contribute most to pricing and how these features are perceived by consumers.

**2. \*\*Data Collection\*\***

- \*\*Data Sources\*\*: Identified and sourced relevant datasets from various platforms, including smartphone specifications, user reviews, and market analyses.

- \*\*Data Types\*\*: Ensured that the data collected included both numeric (e.g., battery life, RAM) and categorical features (e.g., Bluetooth availability, brand).

**3. \*\*Data Preprocessing\*\***

- \*\*Data Cleaning\*\*: Utilized Python libraries (Pandas, NumPy) to clean the dataset by:

- Removing duplicates and irrelevant features.

- Handling missing values through imputation or removal.

- Standardizing data formats (e.g., converting all text to lowercase).

- \*\*Feature Engineering\*\*: Created new features based on existing data to enhance the dataset’s richness (e.g., converting binary values into categorical descriptions).

**4. \*\*Exploratory Data Analysis (EDA)\*\***

- Conducted EDA using visualization tools (e.g., Seaborn, Matplotlib) to:

- Understand data distributions and relationships between features.

- Identify trends, outliers, and potential correlations among features.

- Employed statistical summaries and visualizations (like histograms and box plots) to summarize key findings.

**5. \*\*Feature Selection and Ranking\*\***

- \*\*Correlation Analysis\*\*: Used correlation matrices to identify significant relationships between features and the target variable (price range).

- \*\*Logistic Regression Model\*\*: Developed a logistic regression model to:

- Predict smartphone price categories based on selected features.

- Rank features according to their coefficients, indicating their importance in predicting price.

**6. \*\*Model Evaluation\*\***

**- \*\*Train-Test Split\*\*:** Split the dataset into training and testing sets to validate model performance (75% training, 25% testing).

**- \*\*Performance Metrics\*\*:** Evaluated model performance using metrics such as accuracy, precision, recall, and F1-score to ensure reliability.

**- \*\*Confusion Matrix\*\*:** Analyzed the confusion matrix to gain insights into model predictions and misclassifications.

**7. \*\*Visualization of Results\*\***

- Created comprehensive visualizations using Plotly and Seaborn to present findings:

- Pie charts illustrating the proportion of devices with specific features.

- Distribution plots showcasing feature importance and correlations with price.

- Ensured visualizations were clear and informative for stakeholders.

**8. \*\*Documentation and Reporting\*\***

- Compiled findings, methodologies, and visualizations into a detailed report.

- Prepared a presentation to share results with stakeholders, emphasizing actionable insights derived from the analysis.

**9. \*\*Feedback and Iteration\*\***

- Presented findings to mentors and received constructive feedback.

- Iteratively refined analyses based on feedback, enhancing the robustness of conclusions.

**Assumptions**

**1. \*\*Data Quality and Completeness\*\***

- It is assumed that the datasets used are accurate, reliable, and representative of the smartphone market.

- The data is assumed to be complete, with no significant missing values that could skew the analysis. Any missing values were either imputed or excluded based on their impact on the dataset.

**2. \*\*Feature Importance\*\***

- The assumption is made that the selected features (e.g., battery life, RAM, camera quality) are indeed significant contributors to smartphone pricing and user preferences.

- It is assumed that the features included in the dataset are comprehensive enough to cover all critical aspects that influence a consumer's purchasing decision.

**3. \*\*Consumer Behavior\*\***

- It is assumed that consumer preferences and purchasing decisions are primarily driven by the features included in the dataset.

- It is also assumed that these preferences do not significantly vary across different demographic segments or geographic regions, unless specifically analyzed.

**4. \*\*Market Stability\*\***

- The analysis assumes that the smartphone market conditions (prices, features, competition) are relatively stable during the data collection period.

- It is assumed that the trends observed in the dataset will hold true for future smartphone releases, barring any significant technological advancements or market disruptions.

**5. \*\*Statistical Relationships\*\***

- It is assumed that there is a linear relationship between the selected features and the target variable (price range), allowing the use of logistic regression for analysis.

- The analysis assumes that the relationships observed in the training set will generalize to the test set and to the broader population.

**6. \*\*Normal Distribution of Features\*\***

- It is assumed that many of the features, such as battery power and RAM, follow a normal distribution, which is a common assumption for many statistical analyses.

- The assumption is that the distributions of numerical features do not require complex transformations for the modeling process.

**7. \*\*Feature Independence\*\***

- The assumption is made that the features selected for the model are independent of one another, meaning that one feature's value does not influence another's significantly.

- While multicollinearity may exist, it is assumed that it does not unduly affect the logistic regression model's performance.

**8. \*\*Classification Labels\*\***

- It is assumed that the target variable (price range) is well-defined and that the classification categories are mutually exclusive and collectively exhaustive.

- It is assumed that the labels used for classification accurately represent consumer perceptions of value in the smartphone market.

**Exceptions / Exclusions**

**1. \*\*Data Limitations\*\***

- \*\*Incomplete Data\*\*: Certain smartphone models or features may be underrepresented in the dataset due to incomplete data collection processes, resulting in a potential bias in the analysis.

- \*\*Temporal Limitations\*\*: The analysis is based on a specific time frame; changes in consumer preferences or technological advancements after this period are not accounted for.

**2. \*\*Geographic Scope\*\***

- \*\*Regional Focus\*\*: The analysis primarily focuses on smartphones available in specific geographic regions (e.g., India). Models or features relevant to other regions may not be included, limiting the applicability of the findings to a global context.

- \*\*Market Variability\*\*: Differences in pricing strategies, features, and consumer preferences across different regions are not considered, which may affect the generalizability of the results.

**3. \*\*Exclusion of Certain Features\*\***

- \*\*Feature Selection\*\*: Some features that may influence smartphone pricing and consumer preferences (e.g., brand reputation, marketing strategies) were excluded from the analysis due to data availability or relevance.

- \*\*Secondary Features\*\*: Features such as color, design aesthetics, and brand loyalty were not included, as they are subjective and harder to quantify in the analysis.

**4. \*\*Model Limitations\*\***

- \*\*Methodological Constraints\*\*: The project primarily utilizes logistic regression for classification; other modeling techniques (e.g., decision trees, neural networks) that may provide different insights were not explored.

- \*\*Assumption Violations\*\*: The logistic regression model assumes linearity and independence of features, which may not hold true in all cases, potentially affecting the robustness of the predictions.

**5. \*\*Consumer Behavior Analysis\*\***

- \*\*Lack of Primary Data\*\*: The analysis does not include primary data collection methods, such as surveys or interviews, to understand consumer behavior and preferences directly.

- \*\*Static Analysis\*\*: The analysis does not account for dynamic factors affecting consumer choices, such as promotional events or seasonal trends in smartphone sales.

**6. \*\*Focus on Technical Features\*\***

- \*\*Exclusion of Non-Technical Factors\*\*: Non-technical factors influencing smartphone purchasing decisions, such as social influences or economic conditions, were not analyzed, limiting the scope of the findings to technical specifications only.

**7. \*\*Limited Statistical Generalization\*\***

- \*\*Sample Size Constraints\*\*: The analysis may be limited by the sample size, which can affect the statistical power and generalizability of the conclusions drawn from the dataset.

**8. \*\*Focus on Specific Brands\*\***

- \*\*Brand Exclusions\*\*: Some prominent smartphone brands may not be included in the dataset or analysis, leading to potential biases in understanding market trends.

**Algorithms:**

**1. \*\*Data Preprocessing Algorithms\*\***

**- \*\*StandardScaler\*\***

- **\*\*Description\*\*:** A preprocessing technique used to standardize features by removing the mean and scaling to unit variance. This is particularly useful when features have different units and scales, ensuring that the machine learning model treats each feature equally.

**- \*\*Implementation\*\*:**

from sklearn.preprocessing import StandardScaler

sc\_x = StandardScaler()

x\_scaled = sc\_x.fit\_transform(x)

```

**2. \*\*Feature Selection Algorithms\*\***

- \*\*Recursive Feature Elimination (RFE)\*\*

- \*\*Description\*\*: An algorithm that selects features by recursively considering smaller sets of features. It uses a model (e.g., logistic regression) to identify and rank features based on their importance.

**- \*\*Implementation\*\*:**

from sklearn.feature\_selection import RFE

from sklearn.linear\_model import LogisticRegression

model = LogisticRegression()

rfe = RFE(model, 5) # Selecting top 5 features

fit = rfe.fit(x\_scaled, y)

```

**3. \*\*Classification Algorithms\*\***

- **\*\*Logistic Regression\*\***

- \*\*Description\*\*: A statistical model that uses a logistic function to model a binary dependent variable. It's a popular choice for binary classification problems.

- \*\*Implementation\*\*:

from sklearn.linear\_model import LogisticRegression

model = LogisticRegression()

model.fit(x\_train, y\_train)

predictions = model.predict(x\_test)

```

**- \*\*Decision Tree Classifier\*\***

- \*\*Description\*\*: A non-parametric supervised learning method used for classification and regression tasks. It splits the data into subsets based on the feature values, forming a tree-like structure.

- \*\*Implementation\*\*:

from sklearn.tree import DecisionTreeClassifier

dt\_model = DecisionTreeClassifier()

dt\_model.fit(x\_train, y\_train)

dt\_predictions = dt\_model.predict(x\_test)

```

- **\*\*Random Forest Classifier\*\***

- \*\*Description\*\*: An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification tasks. It helps reduce overfitting.

- \*\*Implementation\*\*:

from sklearn.ensemble import RandomForestClassifier

rf\_model = RandomForestClassifier()

rf\_model.fit(x\_train, y\_train)

rf\_predictions = rf\_model.predict(x\_test)

```

**- \*\*Support Vector Machine (SVM)\*\***

- \*\*Description\*\*: A supervised learning algorithm that can be used for classification or regression challenges. It finds the hyperplane that best divides a dataset into two classes.

- \*\*Implementation\*\*:

from sklearn.svm import SVC

svm\_model = SVC()

svm\_model.fit(x\_train, y\_train)

svm\_predictions = svm\_model.predict(x\_test)

```

**4. \*\*Evaluation Metrics\*\***

- \*\*Confusion Matrix\*\*

- \*\*Description\*\*: A table used to evaluate the performance of a classification model by comparing predicted and actual classifications.

- \*\*Implementation\*\*:

from sklearn.metrics import confusion\_matrix

cm = confusion\_matrix(y\_test, predictions)

```

**- \*\*Classification Report\*\***

- \*\*Description\*\*: A report that provides precision, recall, F1-score, and support for each class in a classification task.

- \*\*Implementation\*\*:

from sklearn.metrics import classification\_report

report = classification\_report(y\_test, predictions)

```

**Challenges & Opportunities**

**Challenges:**

**#### Challenges**

**1. \*\*Data Quality and Completeness\*\***

- \*\*Description\*\*: One of the major challenges faced was ensuring the quality and completeness of the dataset. Missing values, duplicates, and inconsistent data types can lead to biased or inaccurate results.

- \*\*Mitigation\*\*: Implemented data cleaning techniques such as imputation for missing values and normalization to handle inconsistencies.

**2. \*\*Feature Selection and Engineering\*\***

- \*\*Description\*\*: Identifying the most relevant features from a large dataset can be difficult. The initial dataset may contain irrelevant or redundant features that can complicate the model training process.

- \*\*Mitigation\*\*: Employed techniques such as Recursive Feature Elimination (RFE) and correlation matrices to select significant features and enhance model performance.

**3. \*\*Model Overfitting\*\***

- \*\*Description\*\*: Overfitting occurs when a model learns the noise in the training data instead of the underlying patterns, resulting in poor generalization to unseen data.

- \*\*Mitigation\*\*: Utilized techniques like cross-validation, pruning in decision trees, and ensemble methods (e.g., Random Forest) to improve model robustness.

**4. \*\*Interpreting Results\*\***

- \*\*Description\*\*: Interpreting the results of complex models, especially ensemble methods, can be challenging. It may be difficult to communicate the insights gained from the models to stakeholders.

- \*\*Mitigation\*\*: Used model interpretability tools (e.g., SHAP values) to better understand feature importance and model predictions, aiding in clear communication.

**5. \*\*Computational Limitations\*\***

- \*\*Description\*\*: Training complex models on large datasets can be computationally expensive and time-consuming, leading to delays in obtaining results.

- \*\*Mitigation\*\*: Employed dimensionality reduction techniques (e.g., PCA) to reduce the feature space and used cloud computing resources to speed up the training process.

#### Opportunities

**1. \*\*Enhanced Feature Understanding\*\***

- \*\*Opportunity\*\*: The project provides an opportunity to gain insights into the features that significantly impact smartphone pricing and consumer preferences.

- \*\*Action\*\*: Further analysis can be conducted to explore the relationships between features and pricing, allowing for better product design and marketing strategies.

**2. \*\*Implementation of Advanced Models\*\***

- \*\*Opportunity\*\*: There is potential to explore and implement more advanced machine learning algorithms (e.g., Gradient Boosting, Neural Networks) to improve prediction accuracy.

- \*\*Action\*\*: Experimenting with these models can lead to discovering innovative approaches for classifying smartphone features and improving overall performance.

**3. \*\*Deployment of Predictive Models\*\***

- \*\*Opportunity\*\*: The developed models can be deployed in real-world applications, such as recommendation systems or dynamic pricing strategies, for smartphone manufacturers or retailers.

- \*\*Action\*\*: Collaborating with industry partners to integrate the model into their systems could enhance customer experience and increase sales.

**4. \*\*Cross-Domain Applications\*\***

- \*\*Opportunity\*\*: The methodologies developed during the internship can be adapted and applied to other domains, such as automotive, electronics, or consumer goods, for feature ranking and pricing strategies.

- \*\*Action\*\*: Exploring similar datasets from different sectors can yield insights that can inform product development and market positioning.

**5. \*\*Continuous Learning and Improvement\*\***

- \*\*Opportunity\*\*: The project opens avenues for continuous learning by incorporating new data as it becomes available, enabling ongoing model improvement and adaptation to market trends.

- \*\*Action\*\*: Establishing a framework for regular model updates and evaluations can keep the analysis relevant and accurate over time.

**Risk:**

**1. \*\*Data Quality Risks\*\***: Inaccurate or incomplete data can lead to erroneous conclusions and affect model reliability.

**2. \*\*Model Performance Risks\*\*:** The chosen algorithms may not generalize well to unseen data, resulting in poor predictions.

**3. \*\*Implementation Risks\*\***: Integrating models into existing systems may face technical challenges or resistance from stakeholders.

**4. \*\*Compliance Risks\*\*:** Failing to adhere to data privacy regulations could lead to legal issues.

**5. \*\*Resource Risks\*\***: Limited computational resources may hinder model training and analysis, impacting project timelines.

**Recommendations**

In this internship each and every process will explain in the digital discussion room and by our mentor while submitting daily reports . continue the process and give more informations in the digital discussion room it will heip for the Project to develop more…

**Outcome / Conclusion**

The internship provided valuable insights into smartphone feature ranking, showcasing the impact of various specifications on pricing. By leveraging data analysis and machine learning techniques, we developed a model that accurately predicts smartphone price ranges based on their features. This project highlights the importance of data-driven decision-making in the technology sector and lays the foundation for future enhancements in smartphone analytics.

**Enhancement Scope**

future enhancements could include integrating more advanced machine learning algorithms, incorporating real-time market data, and expanding the feature set to include user reviews and brand reputation. Additionally, exploring the impact of emerging technologies, such as AI and 5G, on smartphone pricing and features would further refine the model's accuracy and applicability in a rapidly evolving market.

**Research questions and responses**

Many information will be give in the digital discussion rooms

In this internship I have learned and research many information for browser and YouTube

The link are given below:

Here are some useful resources and links related to smartphone feature ranking, market analysis, and data science methodologies that you can use for your project:

G**eneral Smartphone Market Analysis**

**1. \*\*Statista: Mobile Phone Market Share Worldwide\*\***

[Statista Mobile Phone Market Share](https://www.statista.com/statistics/266136/global-market-share-held-by-smartphone-vendors/)

A comprehensive overview of the smartphone market, including brand shares and trends.

**2. \*\*Gartner: Smartphone Sales and Market Trends\*\***

[Gartner Smartphone Market Share](https://www.gartner.com/en/newsroom/press-releases/2022-02-14-gartner-says-worldwide-smartphone-sales-to-end-users-increased-5-1-percent-in-2021)

Offers insights into sales data and market dynamics.

#**## Feature Analysis and Rankings**

**3. \*\*GSMArena: Mobile Specifications Database\*\***

[GSMArena](https://www.gsmarena.com/)

A detailed database of smartphone specifications, reviews, and comparisons.

**4. \*\*PhoneArena: Smartphone Comparison Tool\*\***

[PhoneArena Comparison Tool](https://www.phonearena.com/phones/compare)

A tool for comparing different smartphone models based on specifications and features.

**### Data Science and Machine Learning Resources**

**5. \*\*Kaggle: Mobile Price Classification Dataset\*\***

[Kaggle Mobile Price Classification Dataset](https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset)

A dataset useful for machine learning projects related to mobile price classification.

**6. \*\*Towards Data Science: Feature Engineering in Machine Learning\*\***

[Feature Engineering in Machine Learning](https://towardsdatascience.com/feature-engineering-in-machine-learning-a-comprehensive-guide-1ab8d98ed7de)

A guide on feature engineering, which is crucial for your project.

**7. \*\*Scikit-learn Documentation\*\***

[Scikit-learn Documentation](https://scikit-learn.org/stable/documentation.html)

Official documentation for Scikit-learn, a Python library for machine learning that you might use for your analysis.

**### Research Papers and Articles**

**8. \*\*ResearchGate: Analysis of Smartphone Features and Pricing\*\***

[ResearchGate](https://www.researchgate.net/)

A platform for finding research articles related to smartphone feature analysis.

**9. \*\*IEEE Xplore: Smartphone Feature Analysis\*\***

[IEEE Xplore](https://ieeexplore.ieee.org/)

Access to various research papers that might be relevant to your topic.

**### Data Visualization Tools**

**10. \*\*Plotly Documentation\*\***

[Plotly Documentation](https://plotly.com/python/)

Official documentation for Plotly, a library for creating interactive graphs and visualizations.