Final Project - Data Science

Title:

Achieving Long-Term Success Through Enhanced Candidate Selection

Highlight:

Qualitative Screening for Advancement:

Based on the answers a candidate provided and the qualities the company identified as significant for the position - the one who passes the threshold proceeds to the second stage.

Cluster Analysis for Long-Term Retention:

Applying the K-Means model to the highest-ranked candidate, dividing into clusters, and identifying a "strong" cluster, characterized by long-term retention.

Identifying Long-Term Candidates:

Correlating the candidate's traits with the likelihood of staying in their job for the long term.

Useful Terminology:

Candidate Profiling:

When hiring for a new position, getting to know the candidate is crucial. Candidate profiling involves gathering and analyzing data about a potential candidate's skills, qualifications, experience, and other attributes. This helps make informed decisions about who to interview and ultimately hire.

Predictive Analytics:

Predictive analytics uses historical data, algorithms, and machine learning to predict future events. In candidate selection, it assesses the likelihood of a candidate's success by analyzing past performance, job fit, and personality traits, leading to better hiring decisions.

Decision Support System:

A decision support system is a computer-based tool that aids decision-making by providing relevant information, analysis, and recommendations. In candidate selection, it offers data-driven insights to help recruiters and hiring managers make more informed and objective decisions.

Optimized Candidate Selection:

Optimized candidate selection improves and refines the methods and criteria for selecting candidates. It uses data, techniques, and tools to make the evaluation process more effective and efficient, helping managers find the right person for the position.

Long-Term Retention Clustering:

Long-term retention clustering helps organizations improve employee retention rates by grouping employees based on characteristics or behaviors related to retention. Identifying these groups provides insights into common patterns, enabling targeted strategies to support long-term retention.

Body:

S.M.A.R.T. Goal:

Specific: Develop a prototype of a candidate selection model incorporating K-Means clustering and qualitative screening to identify candidates likely to stay with the company for the long term.

Measurable: Evaluate the performance of the prototype by conducting a small-scale test with a sample of 50 candidates. Measure the effectiveness of the model by comparing the retention rates of the selected candidate to those selected using the previous selection process. The goal is to achieve a 10% improvement in long-term retention rates based on the test results.

Achievable: Leverage skills and knowledge in data analysis and modeling to design and develop a prototype of the candidate selection model. Utilize open-source libraries and frameworks to streamline the development process and enhance understanding through available online resources and tutorials.

Relevant: Build a prototype of the candidate selection model to demonstrate its potential value to companies by improving their hiring process. This aligns with the common goal of combining computer science and entrepreneurship knowledge to create innovative solutions that address real-world challenges and provide value to organizations.

Time-bound: Develop the prototype of the candidate selection model within three months. Build an MVP, conduct the small-scale test, and evaluate its effectiveness over the following two months to achieve a 10% improvement in long-term retention rates.

Description of Data Preparation + EDA:

In the data preparation and exploratory data analysis phase, several steps were taken to transform and analyze the data for candidate selection.

Firstly, the data was converted into a binary format by assigning a value of 1 to numbers above 0.7 (as 0.7 was found to be a high enough level of confidence to decide if a person has a trait or not) and 0 to numbers below that threshold. This binary conversion helped simplify the data and make it easier to work with in subsequent analyses.

```
threshhold = 0.7
applicants_groups = df.drop(['Question'], axis=1)
applicants_means = applicants_groups.groupby('Applicant').mean().reset_index()
features = applicants_means.drop(['Applicant'], axis=1)
binary_features = features.applymap(lambda x: 1 if x > threshhold else 0)
binary_df = pd.concat([applicants_means['Applicant'], binary_features], axis=1)
```

Executed at 2023.07.14 22:39:53 in 7ms

I created a scoring metric specifically tailored for evaluating candidates. This metric considered various factors and attributes important for the "salesperson" job role (chosen as an example, but the method works with any role). I added the calculated scores as a new column in the dataset, allowing for the ranking and comparison of candidates based on their overall suitability.

During the EDA phase, I identified correlations between different traits and the desired outcomes for the salesperson job position. I explored relevant traits, including being communicative, confident, attentive, resilient, empathetic, problem-solving, action-oriented, and adaptable. I gained insights into which traits were most influential in predicting success in the role.

Modeling:

The modeling process began by utilizing the mean on the data to identify the top candidate who best matched the characteristics required for the salesperson. These characteristics were derived from articles providing insights into the key attributes and qualifications of successful salespeople.

```
target_traits=["communicative","confident","attentive","resilient","empathetic","problem-solving","action-oriented",
 "adaptable"]
binary_features["mean"]=binary_features[target_traits].mean(axis=1)

candidates = 25
top_candidates_df=binary_features.sort_values(by=['mean'],ascending=False).head(candidates)
top_candidates_df_index=top_candidates_df.index
top_candidates_df_index=binary_df.iloc[top_candidates_df_index]
```

By applying the mean calculation, the initial pool of candidates was narrowed down to those who showed the highest levels of compatibility with the job requirements.

Drawing insights from relevant articles and research, a set of key attributes and qualifications crucial for long-term success within the company was established.

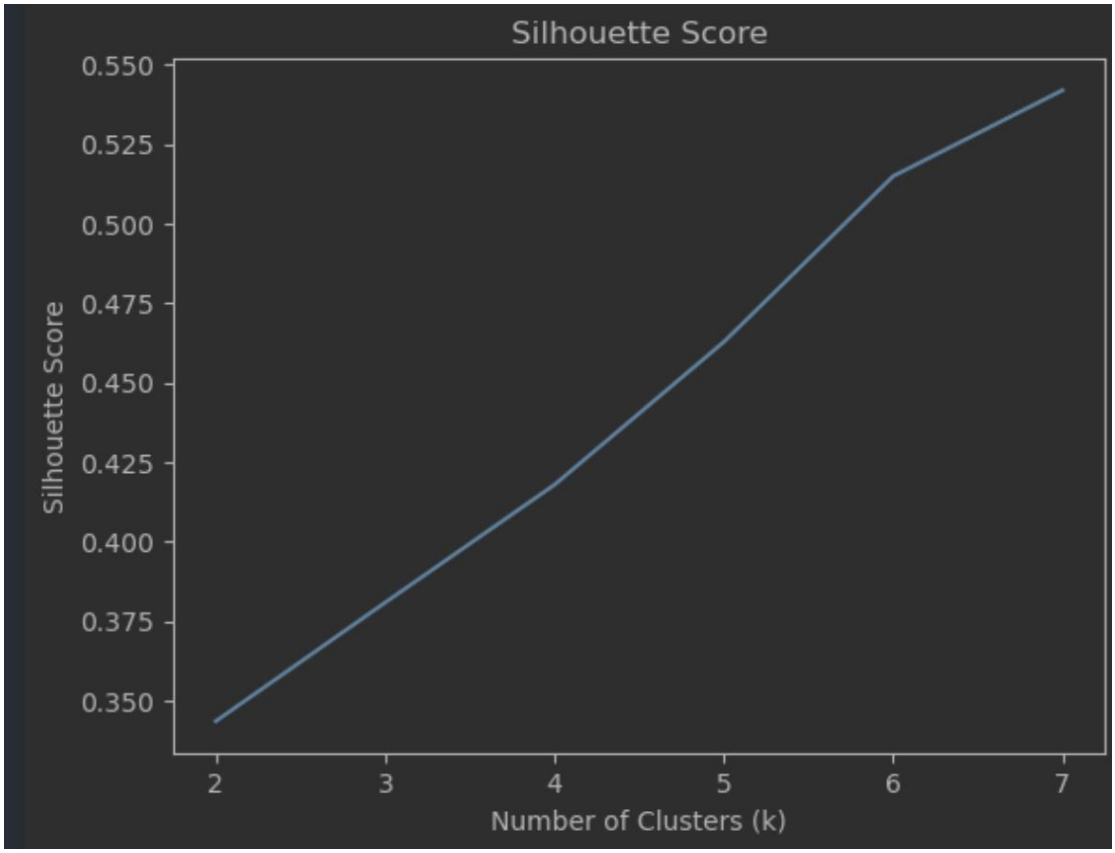
Building upon this initial selection, the K-means clustering algorithm was employed along with the silhouette method to refine the analysis further. The goal was to identify candidates with a higher likelihood of staying in the job long-term. By clustering the candidate based on shared characteristics and using silhouette scores to assess cluster quality, a "strong" cluster characterized by a greater potential for long-term retention was pinpointed.

```python
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import matplotlib.pyplot as plt

X = target_traits_df
silhouette_scores = []
k_values = range(2, 8)  # Range of k values to try

for k in k_values:
    kmeans = KMeans(n_clusters=k)
    kmeans.fit(X)
    labels = kmeans.labels_
    silhouette_scores.append(silhouette_score(X, labels))


# Plotting the silhouette scores
plt.plot(k_values, silhouette_scores)
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Silhouette Score')
plt.title('Silhouette Score')
```

These clustering techniques enabled going beyond individual qualifications and exploring the collective profiles of candidates. By considering various attributes simultaneously, a more comprehensive understanding of which candidates were the best fit for the company and the salesperson position was gained. This data-driven approach prioritized candidates who were not only qualified but also demonstrated a higher probability of staying and growing with the company over the long term, aligning with the goal of enhancing candidate selection for long-term success.

Evaluation:

In evaluating the model, the silhouette score was employed as a metric to assess the quality of the clustering results, as discussed in class. The silhouette score measures how well each candidate fits within their assigned cluster, ranging from -1 to 1, where a higher score indicates a better fit. By considering the silhouette scores, the optimal number of clusters (k) that would yield meaningful and distinct groups of candidates was determined.



The results of the K-means clustering analysis are presented in the form of a multiple bar chart, where each bar represents a cluster identified by the algorithm. By visualizing the clusters and classifying them according to the attributes to be evaluated, informed decisions can be made on which cluster is the best fit for the selection process.

Results:

The results of the K-means clustering analysis are presented in the form of a multiple bar chart, where each bar represents a cluster identified by the algorithm. By visualizing the clusters and classifying them according to the attributes to be evaluated, informed decisions can be made on which cluster is the best fit for the selection process.

Based on the insights gained from the multiple bar chart, the cluster representing the most suitable candidate can be confidently chosen. This candidate not only possesses the necessary qualifications but also exhibits a higher likelihood of staying and thriving within the company, aligning with the goal of enhancing candidate selection for long-term success. This approach ensures that the candidate selection process is optimized for long-term success, increasing the chances of finding an individual who will contribute significantly to the growth and success of the organization.

Conclusions:

- Innervue can improve its candidate selection process by clustering candidates based on job-specific characteristics, such as skills, experience, and personality traits. This will help match the right candidate with the right role and improve long-term success.

- Working on this project as a data scientist has been an eye-opening experience. It has shown the importance of collecting robust data, using advanced analytics techniques, and integrating data into decision-making processes. With this work and ideas, organizations can improve their candidate selection, focusing on achieving long-term success, and make informed decisions that drive their data-driven initiatives forward.

Further Investigation:

- **Candidate Profiling:** What attributes of candidates are most predictive of long-term success, depending on the role?

- **Candidate Screening:** Which screening methods are most effective in identifying candidates who are likely to stay long-term?

- **Talent Analytics:** What are the best practices for using talent analytics to enhance candidate selection and long-term retention?

Bibliography:

1. Barrick, M. R., Mount, M. K., & Gupta, R. (2003). Meta-analysis of the relationship between the Five-Factor Model of personality and Holland's occupational types. Personnel Psychology, 56(1), 45-74.

2. Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. Journal of Applied Psychology, 87(3), 530-541.

3. Allen, D. G., Shore, L. M., & Griffeth, R. W. (2003). The role of perceived organizational support and supportive human resource practices in the turnover process. Journal of Management, 29(1), 99-118.

4. Lewin, J. E., & Sager, J. K. (2010). The Influence of Personal Characteristics and Coping Strategies on Salespersons' Turnover Intentions. Journal of Personal Selling & Sales Management, 30(4), 355-370.

5. Clark, M. (2020). Converting purchase commitments into purchase fulfillments: An examination of salesperson characteristics and influence tactics. Industrial Marketing Management, 85, 97-109.

6. Evans, K. R., Schlacter, J. L., Schultz, R. J., Gremler, D. D., Pass, M., & Wolfe, W. G. (2002). Salesperson and sales manager perceptions of salesperson job characteristics and job outcomes: A perceptual congruence approach. Journal of Marketing Theory & Practice, 10(4), 30.

7. Masputra, H., Nilasari, B. M., & Nisfiannoor, M. (2023). The role of big data predictive analytics as a mediator of the influence of recruitment and selection, remuneration and rewards, training, and development on employee retention. Journal Return, 2(4), 353-365.