County Business Patterns Analysis

By: Sandeep Moparthy, Pooja Kalra,
Janit Modi, Jaebin Park

Presentation Agenda

1 Business Objective & Benefits

2 Data Intro. & Preprocessing

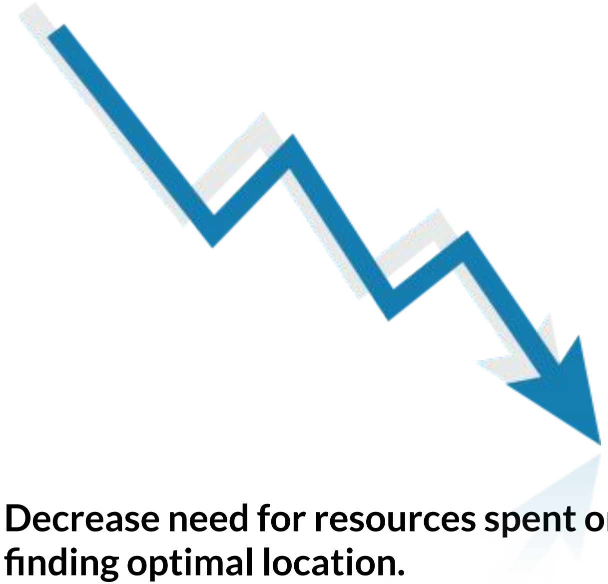
3 Models & Analysis

4 Results

5 Conclusion & Inference

Business Objective

Use data-driven methods to provide target market areas for clients.
Pinpointing the correct target market is a crucial first step towards a successful business plan, leaving the business to move on to the “how”



Decrease need for resources spent on finding optimal location.
As important the location is, the model allows businesses to allocate more time and resources on other areas, such as investment methods.

Three central benefits from the project

A. High, medium, & low potential locations

Provide encompassing results

By taking into account features, such as the underlying industry and employee count in the establishment, we are able to provide results showing both the best and worst areas to start a business

B. Flexibility in Strategizing

Small changes are all that's needed

Though we focused on a smaller portion of the data, all that's needed is a simple change in dataset for more comprehensive and expansive results

C. Provide foundation for new businesses

There are 50 states & thousands of counties.

Whether it's a startup or an established corporation, finding new locations for anything from factories to operations can be difficult. This project derives data driven conclusions to provide the optimal outcome.

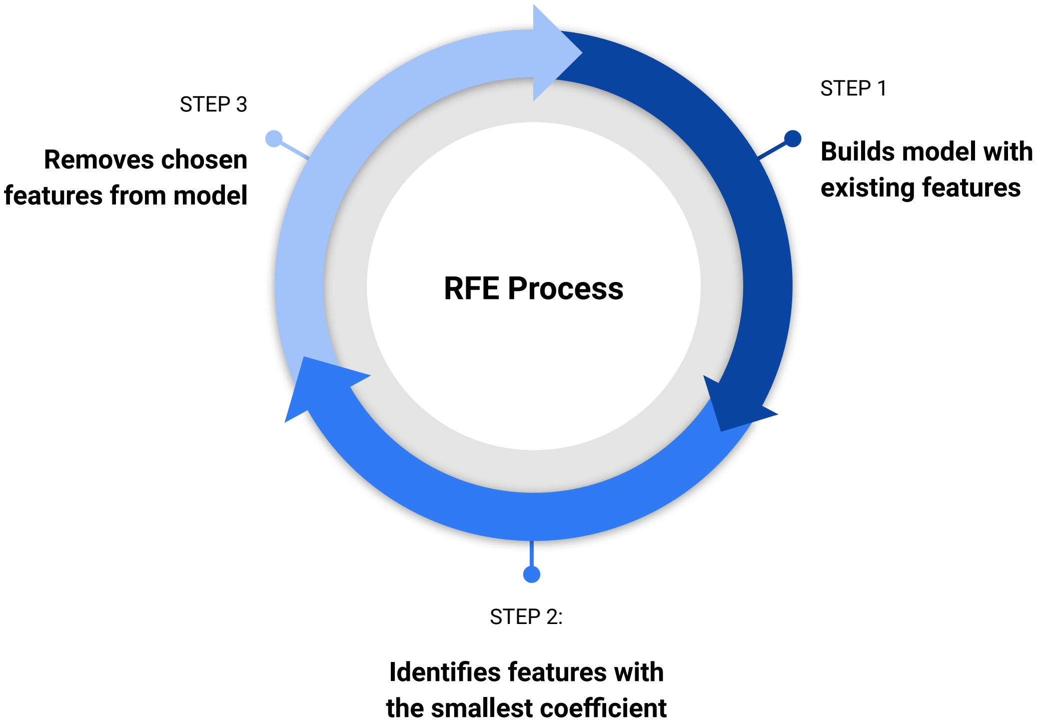
County Business Patterns Dataset

- ➢ County Business Patterns (CBP) → annual series, providing sub-national economic data by industry
- ➢ Levels: US, State, County, Metropolitan, ZIP
  - ○ Classified by NAICS (industry code)
- ➢ Includes 8 datasets, several dozen variables, millions of data points
  - ○ Annual Payroll, Q1 Payroll, # of Establishments, Noise Flags, etc.
- ➢ Drastic need for dimension reduction, data preprocessing,

Data Preparation Process

Drop Unnecessary Variables	Class Size Distinction	NAICS & Merge	State & Industry Selection	Recursive Feature Elimination
From 8 datasets, chose Complete County, US, & State files	Number of Establishments per Industry by Employee Size Class	NAICS had “-” and “/” in it, forcing us to remove them	Dataset still large after merging	Eliminated feature that didn’t strongly contribute to analysis
Removed variables, like Noise Flag, that were unusable in analysis	Approx. dozen columns of size ranges	Attempted to merge with NAICS, but data was too large. Random sampled to *1/sqrt(n)*	Chose to specify Texas, included largest amount of data points	Narrowed down to 22 integral features
	1-99 -> Small 100 - 999 -> Medium 1000+ -> Large	Merged with NAICS	Focused on healthcare industry to replicate client choosing industry	

Recursive Feature Elimination (RFE)



- ➢ Backwards predictor selection method
- ➢ Identifies which variables are important, unlike PCA which shows proportion of variance

Complete List of Variables

- US_lfo_-: US legal form of organization
- US_lfo_G: US legal form of organization Government
- US_lfo_O: US legal form of organization Other
- US_lfo_Z: US legal form of organization S-Corporations

- COUNTY_emp: County total mid-march employees
- COUNTY_q1: C. first quarter payroll
- COUNTY_ap: C. annual payroll
- COUNTY_est: C. total number of establishments
- COUNTY_sz_small: C. establishment class size small
- COUNTY_sz_med: C. establishment class size, medium

- State_lfo_ - : State legal form of organization
- State_lfo_Z: State legal form of organization

- state_q1: State first quarter payroll
- state_ap: S. annual payroll
- state_est: S. total number of establishments
- state_sz_small: S. establishment class size small
- state_sz_med: S. establishment class size, medium
- state_sz_lrg: S. establishment class size, large

- US_sz_small: US establishment class size small
- US_sz_med: US establishment class size, medium
- US_sz_lrg: US establishment class size, large

Fitted Models on AVG. Payroll for Counties

AVG. Payroll For Counties

Classification Methods

- ➢ Logistic Regression
- ➢ Decision Tree
- ➢ SVM
  - ○ Linear, RBF
- ➢ Random Forest
- ➢ Naive Bayes
- ➢ KNN
- ➢ Neural Network

Optimization Method

- ➢ GridsearchCV

Model Evaluation Metrics

- Accuracy - proportion of true results among total number of cases examined

$$\frac{TP + TN}{TP + FP + TN + FN}$$

- Precision - proportion of predicted Positives is truly Positive

$$\frac{TP}{TP + FP}$$

- Recall - proportion of actual Positives is correctly classified

$$\frac{TP}{TP + FN}$$

- F1 - harmonic mean between Precision and Recall, score between 0 and 1

$$2 * \frac{Precision * Recall}{Precision + Recall}$$

Results

Model	Accuracy	Precision / Recall (0)	Precision / Recall (1)	Precision / Recall (2)	Precision / Recall (3)
Log. Reg	0.92	0.96 / 1	0.89 / 0.85	0.68 / 0.56	0.93 / 0.88
RF	0.89	1 / 1	0.88 / 0.85	0.43 / 0.45	0.79 / 0.69
NB	0.80	0.97 / 0.89	0.97 / 0.79	0.41 / 0.43	0.50 / 0.67
KNN	0.42	0.47 / 0.73	0.31 / 0.20	0.00 / 0.00	0.00 / 0.00
SVM - LIN	0.64	0.73 / 0.77	0.59 / 0.59	0.33 / 0.41	1.00 / 0.31
SVM - RBF	0.30	0.46 / 0.35	0.38 / 0.31	0.08 / 0.18	0.07 / 0.12
DT	0.92	1 / 1	0.90 / 0.92	0.63 / 0.55	0.82 / 0.88

- ➢ Portraying models with the best results
- ➢ Classified client data based on the average pay roll into 4 categories:
  - 0 - extremely low
  - 1 - low
  - 2 - medium
  - 3 - high
- ➢ Log. Reg. → best model
  - Accuracy: 0.92
  - Precision Avg: 0.865
  - Recall Avg: 0.823

Conclusions & Inferences: What does it mean?

01	**Adjustable according to client needs**	Depending on the data that's provided, capable of predictions of locations with lowest Avg. Payroll **AND** considerable investment opportunities
03	Enhanced data-driven decision making	Model with 92% accuracy - very little room for error, meaning clients can make lucrative investment decisions without too much risk (location-wise)
04	**Scope for Improvement**	Present model makes predictions only for the healthcare industry in Texas. This can be easily expanded to other states, industries, etc. Possibly expand to develop an interactive and aesthetically pleasing product.