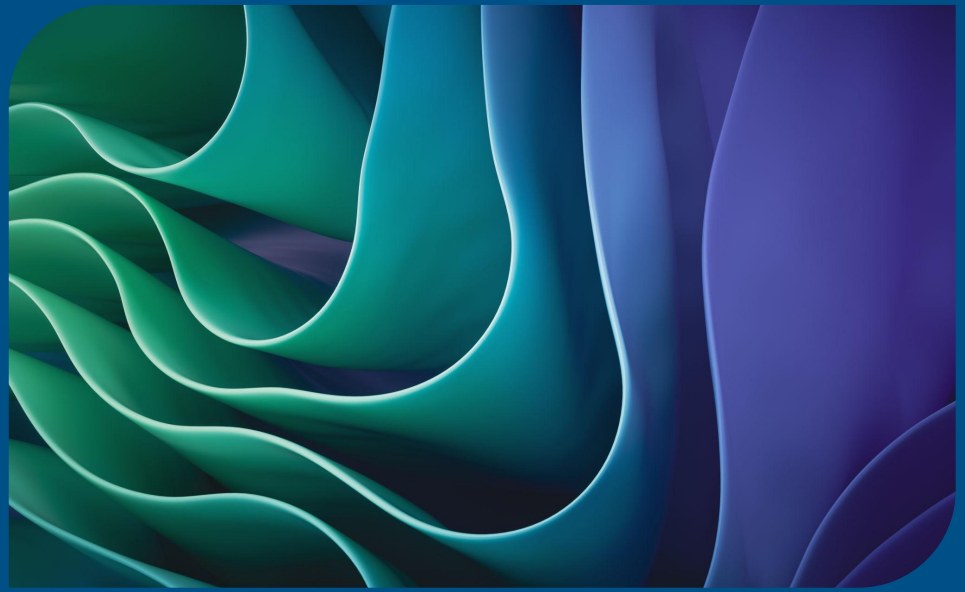
Final Data Analysis Presentation

By: Yisi Lu, Kenny Lei, Alex Fan, Lim Li



Agenda

1. **Business Problem**
2. **About the Data**
3. **Modeling and Analysis**
4. **Dashboard**
5. **Conclusion**

Defining the Problem

"How do household factors such as marriage, childcare responsibility, and household structure affect and influence self-employment status?"

About the Data: Data Source and Variables

- We started by gathering the data covering a ten-year span. We combined these tables and created a dataset with over 3,443,020 records
- CPS Labor Force Characteristics
  - `hhid`, `hhid2`, `lineno`, `married`, `age`, `ch02`, `ch05`, `ch35`, `ch613`, `ch1417`, `hhnum`, `hoh79`, `selfemp`, `selfinc`, `ownchild`
  - The fields are chosen to capture both household-level and individual-level characteristics that may influence self-employment patterns, allowing us to examine how family structure and household composition intersect with self-employment, and provide a comprehensive view of the socio-economic factors at play.

About the Data: Data Cleaning

- Data Cleaning: 3,443,020 records
  - We started off by combining hhid and hhid2. Created a new primary key called hhhid_full
  - We merged selfemp and selfinc as we are only interested in whether or not the person is self employed.
    - Dropped NA's because NA cannot be an integer and we are looking for 0s and 1s
    - Reset the index
  - Ch02, ch35 and ch05 were redundant variables so we removed ch02 and ch35
    - Assumption: A child that is 0-2 years old and 3-5 years old is also 0-5 years old
  - Replaced Null Values with the Mode: ownchild, ch05, ch613, ch1417
  - Filtered the data by age >= 18
    - Assumption: You need to be at least 18 years old to work a proper corporate job without restrictions
  - We converted the fields to numerical data types except the primary key
  - After this process, we were left with 2,212,505 records

Model

Query job e2e80702-c0d8-48d1-b91e-feb660db3524 is DONE. 338.8 MB processed. Open Job

```
Optimization terminated successfully.
         Current function value: 0.320919
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:        self_employment   No. Observations:              2212505
Model:                         Logit   Df Residuals:                  2212496
Method:                          MLE   Df Model:                            8
Date:                Mon, 14 Apr 2025   Pseudo R-squ.:                0.05117
Time:                        00:36:30   Log-Likelihood:            -7.1003e+05
converged:                       True   LL-Null:                   -7.4833e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -4.1139      0.016   -263.298      0.000      -4.145      -4.083
age            0.0410      0.000    221.499      0.000       0.041       0.041
ch05          -0.0120      0.010     -1.155      0.248      -0.032       0.008
married        0.4026      0.008     50.128      0.000       0.387       0.418
hhnum         -0.1464      0.011    -13.395      0.000      -0.168      -0.125
ch613         -0.0517      0.010     -5.158      0.000      -0.071      -0.032
ch1417        -0.1114      0.009    -12.097      0.000      -0.130      -0.093
hoh79         -0.1393      0.009    -15.613      0.000      -0.157      -0.122
ownchild       0.1690      0.005     32.335      0.000       0.159       0.179
==============================================================================
```

```python
# independent and dependent variables
independent_vars = [
    'age', 'ch05', 'married','hhnum',
    'ch613', 'ch1417',
    'hoh79', 'ownchild'
]
dependent_var = 'self_employment'


df = df.to_pandas()

for col in [dependent_var] + independent_vars:
    mode = df[col].mode()
    if not mode.empty:
        df[col] = df[col].fillna(mode[0])


X = df[independent_vars].astype(float)
X = sm.add_constant(X)
y = df[dependent_var].astype(float)


logit_model = sm.Logit(y, X)
result = logit_model.fit()


print(result.summary())
```

Important Takeaways:

- Having Children lowers the rate of self employment
- Older Individuals are more likely to be self employed
- Head of Household is less likely to be self employed
- Owning a child increases the likelihood. Probably if the child is over 18.

Variable	Coefficient	p-value	Interpretation
`const`	-4.1139	0.000	The baseline log-odds of being self-employed when all other variables are 0.
`age`	**0.0410**	0.000	**Significant**. Older individuals are more likely to be self-employed.
`ch05`	-0.0120	0.248	Not significant (p > 0.05). Having children under 5 does **not** significantly affect self-employment.
`married`	**0.4026**	0.000	**Significant**. Married individuals are **more likely** to be self-employed.
`hhnum`	-0.1464	0.000	Significant. Larger household size slightly **decreases** the odds of self-employment.
`ch613`	-0.0517	0.000	Significant. Having children aged 6–13 is **negatively associated** with self-employment.
`ch1417`	-0.1114	0.000	Significant. Same for children aged 14–17.
`hoh79`	-0.1393	0.000	Surprisingly, household head status is **negatively** associated with self-employment here.
`ownchild`	**0.1690**	0.000	Significant. Having any own children **increases** likelihood of self-employment.

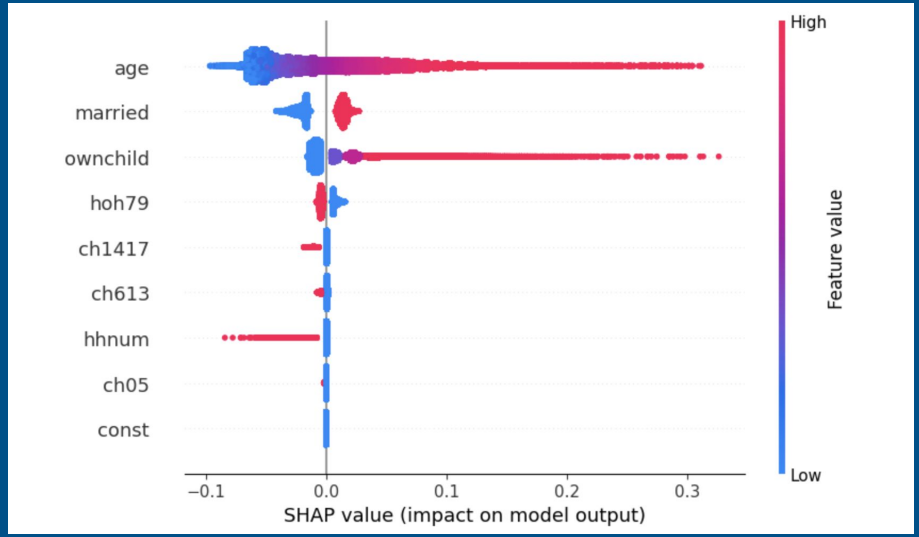
Results of Hypothesis Testing

- Married individuals are more likely to be self-employed than unmarried individuals.
  - Conclusion: Reject the null hypothesis.
  - There is evidence that married individuals are more likely to be self-employed.
- Individuals with children are more likely to pursue self-employment.
  - Conclusion: Reject the null hypothesis.
  - Evidence suggests that individuals with school-aged children are less likely to be self-employed but individuals that own children are more likely .
- Household heads (primary earners) are more likely to be self-employed than other household members.
  - Conclusion: Reject the null hypothesis.
  - There is significant evidence, but in the opposite direction of the hypothesis — household heads are less likely to be self-employed.

Model Analysis

- Age has the strongest influence on the model's prediction
- Married and ownchild are also impactful variables
- Overall, older age positively impacts the prediction outcome. Married status and child ownership also affect predictions but with a more mixed influence.

SHapley Additive exPlanations



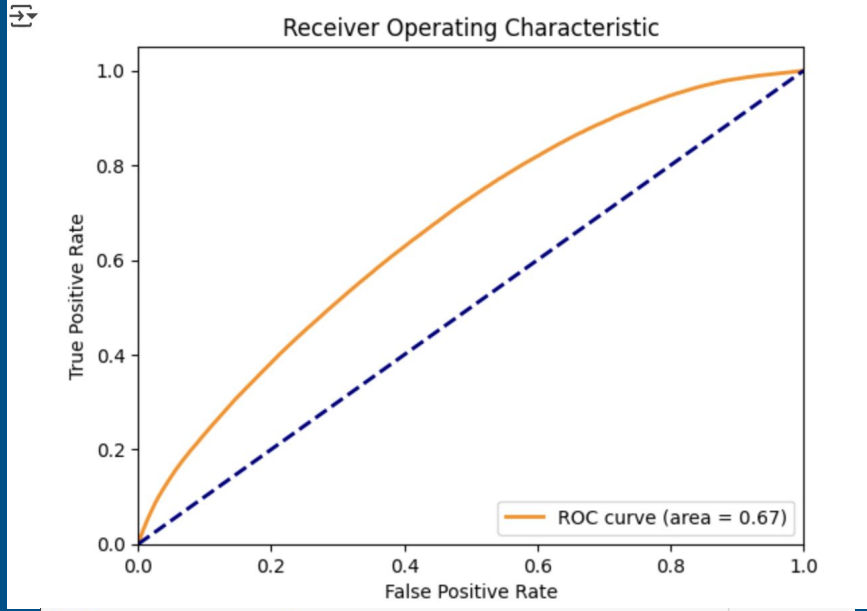

```
# SHAP values
explainer = shap.Explainer(model.predict, X) # Use the model's predict method
shap_values = explainer(X)

# Summary plot
shap.summary_plot(shap_values, X)

# Individual force plot (example for the first observation)
shap.force_plot(explainer.expected_value, shap_values[0,:], X.iloc[0,:])
```

Model Accuracy: ROC

- Our AUC Score is 0.67
- The model has limited ability to distinguish between classes. A perfect model would score 1.0 and a 0.5 would mean random guessing
- Our 0.67 score shows that the model performs slightly better than random but there is room for improvement




```
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, auc
import shap

# Calculate ROC curve and AUC
fpr, tpr, thresholds = roc_curve(y, predicted_probabilities)
roc_auc = auc(fpr, tpr)

# Plot ROC curve
plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=2, label=f'ROC curve (area = {roc_auc:.2f})')
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic')
plt.legend(loc="lower right")
plt.show()
```

Model Accuracy: Confusion Matrix

- Accuracy: 89.39% but this is misleading due to class imbalance
- Class 0 (majority): Precision and recall are both high
- Class 1 (minority): Bad recall of 0.00 meaning the model fails to detect positives
- True negatives 1,977,797
- False positives: 19
- False negatives: 234,683
- True positives: 6
- Overall, this model is biased toward the majority class. It predicts class 1 very poorly making it unsuitable for applications where identifying class 1 is critical

```
Pseudo R-squared (McFadden's R-squared): 0.05117490441009054
Accuracy: 0.8939202397282718
Classification Report:
              precision    recall  f1-score   support

           0       0.89      1.00      0.94   1977816
           1       0.24      0.00      0.00    234689

    accuracy                           0.89   2212505
   macro avg       0.57      0.50      0.47   2212505
weighted avg       0.82      0.89      0.84   2212505

Confusion Matrix:
[[1977797      19]
 [ 234683       6]]
```

```python
# Access the model's accuracy (pseudo-R-squared)
pseudo_r_squared = model.prsquared

print(f"Pseudo R-squared (McFadden's R-squared): {pseudo_r_squared}")

# Get predicted probabilities
predicted_probabilities = model.predict(X)

# Convert probabilities to binary predictions (e.g., using a threshold of 0.5)
predicted_classes = (predicted_probabilities > 0.5).astype(int)

# Compare predicted classes with actual classes
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
accuracy = accuracy_score(y, predicted_classes)
print(f"Accuracy: {accuracy}")

print("Classification Report:")
print(classification_report(y, predicted_classes))

print("Confusion Matrix:")
print(confusion_matrix(y, predicted_classes))
```

Dashboard

Dashboard Insights and Recommendations

- **Context**: We are looking for insights on the characteristics of self-employed individuals in order to provide support, funding, and for policy purposes. We would like to target potential entrepreneurs so that they are offering relevant and efficient services and support to individuals who could become self-employed.

- Based on our model, the SBA should focus on targeting:
  - Older individuals mainly between the ages of
  - Families that own a child over 18 and avoid families with children that are still in school age
  - Married individuals
  - Non head of household members might be multivariate meaning it can have more than one possibility. Head of household might not be a good factor to target.

Drawing a Conclusion

Based on our current model, we suggest that:

- Update the dataset regularly with the most recent demographic and employment trends to account for changes in the workforce, including shifts in marital status, family structure, self-employment, and other variables.
- Incorporate external data sources such as government labor statistics, industry-specific reports, or surveys to provide a broader view of the labor market trends and changes.
- Focus on emerging factors such as the growing influence of the gig economy, remote work trends, and shifts in employee values and preferences regarding work-life balance.
- Examine behavioral data such as job satisfaction, career progression, and aspirations to capture deeper insights into the motivations behind self-employment and its connection to traditional employment roles.

Further analyses should focus on:

- Exploring other significant predictors to see their impact on self-employment.

- Testing different economic conditions or regional variables that may influence self-employment

Deficiencies and Next Steps

- Model is imbalance and not effective for detecting the minority class
  - We can try resampling using Synthetic Minority Oversampling Technique to reduce the chance of oversampling
- The pseudo R-squared is low (around 0.05), indicating that the model explains very little of the variation in the data
- The features may not be strong predictors in their current form or need further processing
  - Create new variables or interactions based on domain knowledge
  - Normalize or scale numerical variables
  - Consider non-linear or multi-colinearity transformations if needed
- Current model is not reliable enough for identifying self-employed individuals — which may affect targeting strategies, policy design, or financial planning.
  - Explore interaction effects between variables (e.g., marital status and children) to understand how these factors combine and influence self-employment decisions. Interaction terms can provide a more nuanced view of the relationships in the labor force.
  - Use advanced modeling techniques like decision trees, random forests, or neural networks to uncover complex relationships and non-linear effects in the data, which logistic regression might not fully capture.

Thank You!